TAILORED TESTING FOR SELECTION AND ALLOCATION.

M.C. KILLCROSS.

Ph.D.

UNIVERSITY OF EDINBURGH.

1975

SUMMARY

The view is proposed that where conditions for psychological testing are favourable the nature of the conventional pencil-and-paper group test sets a limit on the quality of assessment achievable. This is so because it insists on standardisation through uniformity. This procrustean template is a handicap to the assessment of the non-modal man.

Selection and allocation in the Army is a favourable testing situation - centralised, stable and with high volume. Tailored testing is an alternative to conventional testing that allows variation to suit the ability of the person being assessed. The present thesis proposes and tries out a tailored testing procedure aimed at selection and allocation in the Army and other like circumstances.

The research review shows tailored testing to be a post-war interest with statistical antecedents in many non-psychological areas. In the last five years research has grown rapidly, stimulated by the increasing possibilities of online computer-assisted testing.

A tailored testing procedure is proposed that makes few assumptions, makes full use of prior information, conducts a test item by item, and reports its outcome in decision risk terms. The aim is for a coping procedure without critical item requirements.

Real-data simulations are carried out using a large sample of recruits' answers to vocabulary items. An independent conventional verbal test provides a basis for item calibration.

Continued.

The procedure uses two novel indices of item performance concerned with the tails of the empirical item characteristic curves and their interaction with the normative recruit distribution of verbal attainment. These indices are held to be more relevant to effective convergence of the tailoring process.

Good accuracy of convergence is demonstrated by the procedure, and savings in test length for the average recruit as well as greater savings for the non-average.

Empirical studies are needed to investigate the temporal dimension of tailored testing.

ACKNOWLEDGEMENTS

Thanks are due and gladly given to the Army Personnel Research Establishment for its encouragement, to Mr. Dennis McMahon for his wise guidance and support, and especially to my wife and family for their forbearance, understanding, and numerous forms of assistance.

DECLARATION

As required by the regulations I declare that this thesis and the annexed computer programs were written by me and that the research reported is my own. Little of this research has been reported hitherto. Killcross & Cassie (1973) was a general introductory paper for which I was senior author; Killcross (1974) presented a recognisable outline of the proposed tailored testing system. Copies of both papers are in Annex I.

M.C. Killcross.

16 Dec 75

TAILORED TESTING FOR SELECTION AND ALLOCATION.

VOLUME 1

LIST OF FIGURES

LIST OF FIGURES continued

LIST OF TABLES

CONTENTS

CONTENTS continued

CONTENTS continued

All ANNEXES are in Volume 2.

1. BACKGROUND AND AIM

Describing the thoroughness of the revision procedures for the Stanford-Binet scale Terman (1942) writes (p.8), " ..... strange as it may seem there are still clinical psychologists who prefer 'a flexible test that can be adapted to the individual,' one that 'will be custom-made to fit each subject.' Needless to say, the progress of psychometrics has consisted largely in escape from the chaos of subjectivity resulting from the impromptu procedures advocated by the author just quoted." (The author quoted and taken to task by Terman was Kent (1937).)

While not advocating impromptu procedures or a return to the chaos of subjectivity it is the aim of this thesis to propose and demonstrate a method of testing that is flexible, adapted to the individual testee, and custom-made to fit him.

Anastasi (1954) writes (pp. 22-24), "A psychological test is essentially an objective and standardised measure of a sample of behaviour." ...., "Standardisation implies uniformity of procedure in administering and scoring the test. ..... Such a requirement is only a special application of the need for controlled conditions in all scientific observations. In a test situation, the single independent variable is the individual being tested." ..... "Such standardisation extends to the exact materials employed, time limits, ... and every other detail of the testing situation."

Both authorities speak without reservation of the virtue of uniformity. The achievement of an unvarying presentation is seen as providing an armour by which a test might more strongly resist the thrust of extraneous influence. These views could be put so emphatically because they were given in a period when psychologists

were comforted and perhaps in part beguiled by demonstrations that their youthful science had its own examples of scientific rigour.

What had happened was that uniformity had been imposed under the umbrella of standardisation. This is one way to achieve standardisation - and at the time was probably the best way - but it is not the only way.

DuBois (1970) considers the test item a very remarkable invention and traces its development from a crude subjective form in 1902 to an item-analysis supported entity that by the US Army Alpha test of World War 1 had already acquired some of what was to be its considerable sophistication. The emphasis came to be on group testing and on the test as a homogeneous assembly of items. That the assembly was also uniform for all testees was not merely a fashionable display of rigour: the undoubted philosophical congeniality of the requirement was secondary to its technical merit. The forceful advance of psychological measurement during the period when called on by the needs of two world wars was critically assisted by the benefits of cultivated uniformity of presentation. Nonetheless, although such uniformity contained a psychometric truth this was not the whole truth. That uniformity was a constraint caused little concern and attracted little attention or even realisation. Indeed variability was generally equated with loss of adequate control - hence Terman's chiding. Yet the general philosophy of psychological measurement is to sample behaviour with a view to extrapolating to a wider area of behaviour about which the sample is informative. So the choice of behaviour sample is critical. If it is now stipulated that an estimate of intelligence, say, must be achieved from the same sample of behaviour from all <u>irrespective of their ability</u> then this can be

seen to be a handicap. Why ask a lengthy fixed series of questions of which perhaps a half are likely to be dead wood, being either too hard or too easy to do any useful work? Would not questions adjusted to ability provide a better sample? Yet the handicap of uniformity was not only accepted but advocated.

Why was the constraint of uniformity accepted? One answer is that removal of the constraint appears to depend on knowing the result of the test before it is given - on knowing the outcome before picking the questions. This answer, however, assumes that a test is an indivisible unit, whereas testing can alternatively be seen as an ongoing process in which it is possible to intervene. A test from this view-point is a series of encounters between testee and test item, each encounter providing a little more information and improving an accumulating estimate of the characteristic being measured. Why then was testing mostly assumed to be a unit and not viewed as a process? One reason was perhaps an incomplete emancipation from pre-Darwinian physicalism; while individual differences and dynamic change were central to the Darwinian theme and directly contributory to later interest in psychological measurement, the traditional methodologies of the physical sciences - suited to the manipulation of single passive variables and to total repeatability - continued to be influential. Among the physicalist concepts were the ideas that a measurement was something extracted in its entirety from a relatively defined situation, and that a measuring instrument was an enduring piece of equipment used for a particular variable over a range of values.

The early individual intelligence tests did acknowledge the redundancy of administering test questions which were too easy or

too hard, and within the overall sequence of the test starting and stopping places were matched to the individual testee: even some choice in the ordering and selection of sub-tests was left to the tester. Perhaps the earlier test constructors were less blinkered than their successors became, or perhaps their flexibility simply represented an unresolved vagueness in conceptualisation. In either case later events moved strongly towards the elimination of any residual variability, not only for large scale testing but also for subsequent individual tests.

For large scale testing the constraint of uniformity also had a sound practical reason, the lack of an alternative flexible technology. The pencil-and-paper testing medium developed to facilitate large scale testing was a solution dictated by the technology of the day and made group testing synonymous with a uniform treatment of the group. For such group testing the move to strict uniformity to ensure truly common treatment is necessary, and one of the consequences of this has been an unproductive spill-over of the philosophy to other measurement approaches.

From as early as the late 1940s the idea of a variable test has been mooted. Variable in the sense that the test is deliberately varied to suit the individual testee, and moreover varied dynamically during testing using previous answers to help select later questions. One form of assessment has indeed always followed an individually varied course, and this is the interview: in a sense the last sentence describes an interview. In a sense too the aim of a variable test can be construed as trying to keep the reliability of a conventional pencil-and-paper test while giving it something of the individuality (and humanity) of the interview. It was hardly an accident that Terman was remonstrating above with a clinical psychologist

- a specialism in which the needs for assessment and recognition of individuality come together. Hutt (1947), also writing from a clinical setting, used the term adaptive testing to refer to an individually adjusted method of Stanford-Binet administration. However, the slant of this thesis is not clinical; it is concerned rather with objective forms of individualised testing.

The development of the idea of testing as an ongoing process adjusted to the individual testee will be presented in the next chapter, but in general terms the aim of such adjustment is greater efficiency and more recently this aim has perhaps been joined by one of greater individual consideration. In a very forward looking paper Hick (1950), considering intelligence tests in the light of information theory, puts forward in embryo an outline of an individually adjusted test (p.161), "Hence an intelligence test should, in theory, be a 'branch process'; i.e. the first question should have a 0·5 chance of being answered by anyone from the general population. If the subject answers it, the next should have a 0·5 chance of being answered by anyone who has been successful with the first; and so on." Hick is considering here only the information transmission aspects of efficiency.

The term tailored testing was coined by Lord (at a 1968 conference in a paper subsequently published as Lord (1970b)). This is the generic name adopted here for all forms of individualised testing. Traditional forms of pencil-and-paper tests will generally be referred to as conventional tests.

Progress in computer technology has now provided a practical medium for tailored testing. An individually tailored test can be conveniently given by maintaining a pool of test questions in computer

storage and sitting the testee at a linked terminal which has a visual display unit (VDU) - of television screen type - and a keyboard. Questions are presented on the VDU and the testee answers on the keyboard. Depending on his performance on earlier questions the testee's next question is chosen by computer programme to match a running estimate (of his ability for example) that is being up-dated as testing proceeds. The attempt is made to optimise the questions chosen for presentation. The same technology has enjoyed wider exploitation within Computer-Assisted Instruction. It is this technology that is in mind for the Army context of this thesis. Tailored testing approaches have also been tried using computer teletype terminals and by various non-computer methods.

Although a convenient technological solution is necessary for any application of tailored testing, the main difficulty has been rather the development of an effective conceptual framework. In an individualised test different testees will take different questions. This immediately removes a cornerstone of classical psychometric theory and creates numerous problems of how to place testees on a common scale. The conceptual hiatus for early researchers in tailored testing should not be underestimated. Lord (1971e) - introducing tailored testing to a statistical readership - felt able to write (p.707), "However, the statistical reader need not be familiar with [even] the basic ideas of classical mental test theory - in particular, the notions of 'true score' and 'reliability'. The in-consequence of the classical theory here is surprising. Perhaps this indicates that the approach to be used is no less fundamental than the classical theory itself."

This thesis puts forward a conceptual framework which owes much

to earlier researchers but which includes some novel points. A method of tailored testing is proposed which is seen as having some advantages. Although of potentially wider scope the research to be described is aimed particularly at application in selection and allocation in the British Army. A description follows of the main points of the present Army entry procedures.

Non-commissioned entry into the Army for men and juniors is a two stage procedure. There are some variations in detail as between men and juniors, and as between men in Scotland and men from the rest of the United Kingdom: in principle, however, the broad approach is the same, and WRAC entry also has moved towards increased conformity with this general framework. The largest applicant group following one particular variant of the general scheme is men in England, Wales and Northern Ireland. This group provides - as will be described - the most fertile ground for the introduction of individualised testing, and the details and discussion which follow refer to this context and not necessarily to the other procedural variants.

The first stage of the entry procedure takes place at the local Army Careers Information Office (ACIO). This is the screening stage at which both applicants and Army take most of their decisions about broad suitability. The Army's screening decision is based on biographical and educational information from interview, on the results of a 30 minute pencil-and-paper test of general reasoning ability and basic arithmetic and verbal attainment, and on a medical examination. About 50% of applicants are screened out at the ACIOs.

Successful applicants go on to the Recruit Selection Centre (RSC) at Sutton Coldfield for the 2 -2½ days which make up the second stage of the entry procedure. The men who go on to RSC are technically

recuits, having enlisted at the ACIO; however, this enlistment is not binding and an honourable discharge is freely available whilst at RSC. Depending on recruitment between 12,000 and 20,000 recruits may pass through RSC in a year and only about 10% drop out from all causes.

At RSC the concern is mainly with allocation, with finding the best match between the interests and abilities of the recruit and the needs of the Army. This matching process is helped by a two-way information flow: the recruit learns about the Army and about the employments available and the Army learns more about the recruit. For the recruit a formal and extensive job briefing is supplemented by a question and answer session and by interview and informal discussion opportunities; for the Army cognitive abilities and attainments are measured by a standard set of five pencil-and-paper tests, occupational interests and motivation are assessed by a pencil-and-paper inventory and at interview, and a detailed medical examination is carried out. At interview a Personnel Selection Officer (PSO) continues the information exchange and ultimately helps the recruit decide on his first three allocation preferences. One of these is almost always offered to the recruit, and in 75% of cases it is his first choice. Now the recruit must accept the allocation or claim his discharge. With the minimum of delay the allocated recruit will be posted to his training depot.

The Army's selection and allocation procedures are seen to follow the traditional pattern based on standardised pencil-and-paper tests designed for uniform group administration. Up to five years ago this pattern was more appropriate, but then the present two-stage centralised procedures were introduced. Previously selection and initial allocation had both taken place at the many ACIOs, and the fuller

pencil-and-paper testing with the five standard tests had been carried out only at a recruit's allocated training depot - where his final allocation was generally confirmed within the small range of employments available there. This was the traditional context calling for uniform test treatment. With many testers in many testing locations a system with a heavy emphasis on uniformity is exactly what is needed to combat the wide variety of experience that an allocated recruit would encounter. The traditional group test could be relied upon to give meaningful results when given by almost anyone almost anywhere - or, more moderately, administered according to its rules it was robust in resisting the potential influence of rather wide background variations.

It can be speculated that computer-linked remote terminals might have been possible at ACIOs or training depots even within an uncentralised framework, but clearly many more terminals would be needed and their usage would be less intensive than if installed at an all-Army Selection Centre. But now that there has been the change-over to centralised selection, and now that an alternative flexible technology is available, we are left with the paradox that the earlier necessary emphasis on uniformity has become an overkill. What was a protective armour is now an impediment, what was so successful in safe-guarding a minimum standard now limits the maximum. In evolutionary terms the earlier adaptation based on a standard selection test battery has become maladaptive for the new environment. It is as if all recruits continued to be issued with one size of uniform even though tailoring capacity has become available.

The thesis presented here is that benefits are to be expected from a move to an individualised approach to psychological measurement:

and in particular that cognitive testing for selection and allocation in the Army at high volume centres is well placed to benefit from the improved behaviour sampling that individualised measurement allows. The idea of tailored testing is limited here to the measurement of one cognitive characteristic at a time. The logical extension of tailored testing to varying the set of characteristics assessed is not pursued - a standard set of tests limits the behaviour samples available which would manifestly gain if suited to an individual recruit's job probabilities.

Tailored testing involves some kind of item selection from an item pool. The item is becoming the working unit rather than the whole test. It is perhaps not entirely fanciful to see an analogy between the move from the test to the individual question, and the progress of physical science through successive layers of increasingly microscopic levels of inquiry. Similarly from a wider view one can see parallels between this move to individualisation and the reaction against mass production, conformity, and increasing individual anonymity. It may be in the end that individualised procedures will be adopted simply because they are individual and not because of technical superiority.

2. REVIEW OF RESEARCH AND INTRODUCTORY DISCUSSION.

Research on tailored testing is a post-war phenomenon. Even so, despite having a history of 25 to 30 years behind it, such research has not grown to any great volume. Tailored testing has been a persisting idea, cherished in turn by a series of researchers, but laid down almost as often as taken up. It has held a promise which it has been slow to fulfil. Wood (1973) speaking of educational interest in tailored testing comments (p. 529), "For the past twenty-five years, this idea has exerted a more or less continuous fascination on the educational research community, and there has probably always been someone working on it. Yet with the greatest respect to all concerned, these enquiries have never really amounted to anything of practical significance." The attraction of the idea is almost tangible but an operational real-life application has yet to emerge. The traditional pencil-and-paper group test is, of course, one of psychology's major successes. It tends to monopolise educational and psychological assessment as IBM and Hoover have monopolised other fields. The major traditional test users have developed smooth-running, effective procedures that they are unlikely to be persuaded from by merely modest temptation. However, research on tailored testing is now growing: most of the literature references made in this review to work specifically on tailored testing were published after 1970. The following table illustrates the growth.

Table 1. The distribution over time of published literature on research in tailored testing.

Period	No. of publications
1944 - 47	2
1948 - 51	2
1952 - 55	2
1956 - 59	2
1960 - 63	3
1964 - 67	6
1968 - 71	21
1972 - 75	22

The growth can be linked to the availability of a facilitating technology - that of time-shared, fast computers. The earliest experimental studies - up to the late '60's in some cases - perforce attempted pencil-and-paper implementations of tailored testing. This was a stony road and a testimony to the drive and ingenuity of the researchers; such administrative inconvenience weighed heavily against any subsequent application. Much of the earlier literature also tended to be conceptual and concerned with theoretical results. The use of an on-line terminal was quickly seen as almost a pre-condition of a workable tailored testing system, and research gives every appearance of marking time while the technology ripened. It may be too that the upturn of concern for the individual characteristic of recent years has provided a more supportive climate.

The nature of research on a topic develops and matures, but not evenly in different countries or across professional specialisms. Ideas persist with some groups, are dropped quickly by others or are perhaps never taken up at all. The analogy of changes in fashions

of dress would have several points of correspondence. Consequently, although this review will parcel up the literature into four bundles, little more is claimed for the classification than a certain structural and conceptual usefulness.

The research literature may be grouped as follows:-

A. Statistical antecedents.

B. Effects of item context.

C. Background tailored testing research.

D. Recent tailored testing research.

A and C are important for their conceptual contributions rather than their detailed findings. C refers mainly to early pre-computer work or research peripheral to this thesis. B is necessary to establish the case that items can in some circumstances be considered as independent units. D contains much detail of value and constitutes the main substance of the review: the writer will attempt to show that the ideas on tailored testing he presents later are consistent with experience so far and represent a promising approach. The approach to be adopted (described in Chapter 3) builds on a number of pointers from this literature but also includes a number of novel features. In this latter sense the existing literature does not include a direct ancestral line.

A. Statistical Antecedents

In this background section the aim is to trace the development of statistical methods that have provided the bases for a variety of approaches to tailored testing. These methods have usually originated with a view to applications in assessment or estimation problems outside psychological measurement. The main distinguishing feature of the methods is that they call for a sequential approach. Such

an approach does not specify a one-piece experiment to be carried through in toto in order to permit estimation of the parameter of interest, rather it proceeds by a sequence of trials. These trials are not pre-determined, instead the specification for each trial is dependent upon the results of the preceding trial sequence. A second distinguishing feature is the type of data to which the methods may be applied. These methods are concerned with dichotomous experimental responses (in our case wrong or right answers to test questions) - usually referred to as quantal response data.

There are two threads to be followed through the development of sequential methods of estimation. The one which emerges as of less persisting interest is that associated with the Statistical Research Group of Columbia University (1945) and Wald (1947, 1950). This approach, sequential analysis, has been adopted extensively in the quality control procedures of manufacturing industries and is applicable where there is a large number of ostensibly equivalent items (rivets, resistors, spools of thread and the like). The problem here is how to sample effectively so as to estimate the level of a characteristic in a particular batch of output. (In our case we are wanting to estimate the ability or attainment of a person (the batch) from his responses to a sample of questions.) The sequential analysis solution is to take items one at a time and check if each in turn meets the required quality standard - thus providing a stream of yes/no data. After an item has been examined the additional evidence is used to update an appropriate cumulative statistic - for example, the Sequential Probability Ratio (Wald, 1947). Depending on the new value of the statistic a decision is taken either to classify the batch finally as acceptable or unacceptable or to

increase the sample by taking in a further item, in which case the sequential procedure is repeated. This final classification decision is made with prescribed risks of false-rejection and false-acceptance. The tailored element of sequential analysis is thus the length of the sequence. The procedure concerns itself with successive decisions about whether there should be a next item or not, there is no question of tailoring the nature of the item. In psychological measurement it would not usually be appropriate or often possible to present a series of test questions that could be regarded as identical in nature. If it were sufficiently certain that such a series was appropriate this would in many cases mean that a sufficient estimate was already available. However, researchers, from Cowden (1946) to Ferguson (1971), have used this approach either as a first approximation or in an educational setting for mastery testing. In the latter case, and especially in relation to criterion-referenced testing where a specific accomplishment is involved, it can well be a matter of repetitive testing (say, of division of fractions) to establish whether acceptable proficiency has been achieved. Research using sequential analysis in individualised educational and psychological measurement is reviewed in Section C. A statistical development by Armitage (1950) supports a multiple final classification rather than a simple split and is used by one group of researches to be described.

Sequential analysis, then, forms the basis of useful but limited applications of tailored testing. It is not a method applicable to the general measurement problem where there is no fixed value in mind for the parameter being estimated. It is not the method followed in this thesis. However, the explicit formulation of decision risks is a characteristic relevant to the tailored testing application in

view and discussion of this feature will form part of the tailored testing method to be proposed. It will emerge that the second thread to be picked up from the development of sequential methods, while generally more helpful, does not have a decision risk orientation.

The second thread leads to the research of most direct relevance to this thesis. It began with methods originally devised for testing the sensitivity of explosives by Anderson, McCarthy and Tukey (1946) of the Statistical Research Group at Princeton. These are the "staircase" or "up-and-down" methods of sequential estimation. Dixon and Mood (1948) suggested that these methods could be applied in other fields and proposed estimators that were taken up in bioassay or toxicology. These ideas were subsequently developed extensively in bioassay, and much of this work offers useful comparison with its psychometric equivalent. Lord (1970 b) in what amounts to a foundation contribution to much recent tailored testing research writes, (p. 140), "It is a fortunate fact that most of the problems dealt with here closely parallel similar problems in bioassay. Much fruitful work has been done on the bioassay problems. This provides the inspiration, the background, and indeed the backbone of this chapter."

It may be helpful to look at the analogy between bioassay and psychometrics in a little detail before following developments further. The bioassayist has an insecticide, say, for which he is trying to estimate the lethality. He has control over the dose administered and can observe death or survival in his insects. The confrontation between dose and insect results in a quantal outcome, life or death. This is analogous to the confrontation between a person's ability and a test question and the outcome fail or pass. The analogy does not hold for what is controlled. Whereas we have information about the

difficulty of our test questions and try to infer an unknown ability by manipulating question difficulty, the bioassayist is unable to vary the resistance of his insects and infers the lethality of his insecticide through varying its dose.

The essence of up-and-down methods is that after a trial with an observed outcome the independent variable is altered for the next trial so as to favour the opposite outcome - after survival the dose is increased, after a wrong answer an easier question is asked. In this way an overall balance of outcomes tends to be achieved. The details of up-and-down procedures are concerned with how the next trial is to be specified (for example, in what steps should the independent variable be changed), with how a decision to terminate the trials is to be made, and with how the observed responses are to be converted to a final estimate or score.

Further developments of up-and-down methods in bioassay are outlined below. A few additional details will also be given in Section D, where they can be more appropriately mentioned after the introduction of background theory which it would be unhelpful to present here in a general statistical context. Brownlee, Hodges and Rosenblatt (1953) wrote of the slow initial take-up of Dixon and Mood's (1948) proposals (p. 262), "In spite of this efficiency advantage, the up-and-down method does not seem to have been given much consideration in such fields as bioassay or fatigue testing of metals." They went on to confirm the superiority of the sequential approach over previous probit methods even for small samples, and proposed the use of a more convenient estimator of the main parameter of interest. They also proposed the possible use of two (or more) parallel series of trials. In their case they were concerned to make good use of the delay some-

times necessary between consecutive bioassay trials, but this idea is of some interest in tailored testing as a means of checking possible anomalous responses and will be brought up for discussion again.

The block up-and-down method is a straightforward extension, of convenience in bioassay, that treats several insects in one trial. This convenience does not translate to tailored testing but administering blocks of questions does permit more complicated rules for choosing the next block and such approaches have attracted some tailored testing research. In bioassay the blocking method has been investigated by Wetherill (1963), Cochran and Davis (1964), and Tsutakawa (1967).

A different development is that by Robbins and Monro (1951). Their proposal may be regarded as a shrinking-step up-and-down method. Larger alterations are made to the independent variable initially, with ever smaller steps as the procedure zeroes in on the appropriate level. Such methods were found by Wetherill (1963) to be extremely satisfactory in some instances. It will be seen later that full Robbins-Monro approaches are not possible in tailored testing, but modified shrinking-step procedures have been proposed. The Robbins-Monro proposals were for large samples. Cochran and Davis (1965) investigated a number of Robbins-Monro procedures for samples of fifty and less and were able to offer useful gains over non-sequential designs. Davis (1971) compared several sequential bioassay methods and concluded that delayed variants of both Robbins-Monro and up-and-down procedures gave good results in all situations. Writing so recently he was, however, still able to comment (p. 80), "While the asymptotic properties of sequential experiments, especially the Robbins-Monro process, are relatively well established, the accuracy of estimates

and the guiding principles for the design of small sample experiments in bioassay are as yet incompletely explored."

A new approach, and one paralleled in tailored testing at about the same time, is that of Freeman (1970) who introduces Bayesian sequential estimation. This will be developed in Section D but it is of interest that research in tailored testing appears at about this stage to be coming abreast of general advances in stochastic approximation.

Before leaving this section it is appropriate to mention for completeness that research in psychophysics has also taken an interest in the developments in sequential estimation in other fields. Cornsweet (1962), Taylor and Creelman (1967) with their Parametric Estimation by Sequential Testing, Kappauf (1969) and Rose et al (1970) are examples. As the psycho-physicist (in common with the bioassayist) is in a position to vary the physical intensity of his stimuli (the analogue here of ability) rather than the sensitivity of his subjects there has not apparently, perhaps for this reason, been any effective cross-fertilisation with tailored testing.

B. Effects of Item Context

Tailored testing methods select for presentation questions drawn from a larger pool. Two testees may well receive no questions in common; when they do receive the same question it will in most cases follow different preceding items and occur at a different stage in their test session. Does such variation in context affect an item's psychometric characteristics? The tailored testing procedures proposed have all assumed that item characteristics will remain stable irrespective of context. It will be necessary for real-life applications of tailored

testing to examine the size of any context effects. In the meantime to maintain the viability of tailored testing research it is sufficient to establish the possibility of context-free items. To establish this possibility, or even probability, is the aim of this Section.

"Question" and "item" have been used interchangeably in the previous paragraph and will be used in this way throughout. While "item" and "response" are strictly more accurate - because many tests are not in interrogative form - "question" and "answer" often allow a less stilted description and are used here with a more general meaning than literally theirs.

It is a priori likely that the content and difficulty of a question series could be made such as to influence the performance of some constituent items. However, the appropriate question is not, "Are substantial effects possible?" but rather, "Are effects likely?" A tailored test tries to present homogeneous items of about the same difficulty. It is in the nature of tailored tests that item difficulty is concentrated in a more or less narrow band appropriate for the testee. In this way the individualised approach might avoid the worst situations for effects stemming from frustation or demotivation.

Investigations that have been made of context effects have been confined to pencil-and-paper tests. The use of a Visual Display Unit (VDU) computer terminal as the testing medium is a change that demands caution when looking to findings from pencil-and-paper settings. Accordingly the research findings reviewed below cannot be taken as definitive: their function is simply supportive. All the studies reviewed were carried out in schools or colleges. (For this reason Chapter 4 will include some supporting evidence from an Army sample.)

Mollenkopf (1950), Sax and Cromack (1966), and Flaugher, Melton and Myers (1968) establish the basic general finding that under essentially non-speeded conditions item statistics and correlations with other variables are not significantly affected by item rearrangement. Sax and Cromack conclude (p. 311), "In general, the results support the thesis that test constructors have a responsibility of arranging items in ascending order of difficulty if tests are lengthy or time limits restricted. Evidently, little is gained in arranging items if time limits are generous. Nor is there any advantage in constructing 'motivational' tests consisting of a few easy items mixed with more difficult ones over random forms of item arrangements."

Marso (1970) carried out two experiments. In the first a pool of two hundred 4-option multiple-choice vocabulary items was used to assemble a 139-item test displaying a wide range of difficulty. This test was arranged in three formats,

- ascending order of difficulty
- descending order of difficulty
- randomly arranged.

The three forms were randomly assigned to one hundred and twenty two students, previously classified as high, average, or low on test anxiety, and administered as power tests. The different item arrangements were found not to relate to score achieved. Test anxiety did affect achievement score but did not interact with item arrangement.

In a second experiment Marso used a course examination arranged again in three forms,

- topic presentation in course order
- topic presentation in reverse order to that of the course

- questions randomly arranged.

Results confirmed those of the first experiment.

A number of studies have looked at item context in the course of investigations of item sampling for estimating test norms (Lord (1962), (1965)). Here subsets of items are administered and used to estimate the mean and standard deviation of the whole test. Any systematic effects on item performance would evidence themselves in systematic errors of estimation.

Owens and Stufflebeam (1969) comparing contrasting samples of about two thousand 4th grade schoolchildren used item subsets of 3,6, and 9 items from 50 multiple-choice vocabulary questions. 17, 8, and 4 different subsets of these three lengths respectively were administered to fractions of each sample. The population mean and standard deviation were as well estimated from item samples as from equivalent pupil samples. Both sampling techniques showed less precision in estimates of the mean for the higher ability pupils from advantaged neighbourhoods. Pupils, having taken an item subset, went on to attempt the rest of the 50 items, so that Owens and Stufflebeam were also able to look specifically at whether variations of item sequence affected test performance. The results from varied and standard sequences were so close as to suggest, the authors conclude cautiously, "..... that the sequence of items need not have a significant effect on test performance," (p. 82).

Sirotnik (1970) looked specifically at the context effect in item sampling. He investigated mean and variance estimates from subsets of vocabulary (synonym), arithmetic, and teacher attitude items taken by 180 students under power conditions. No support for

a context effect was found, the author giving his opinion that the mean estimates were relatively immune to context effect for all three types of item, while further studies were needed to look at variance estimation.

Feldt and Forsyth (1974) looked at the same topic as Sirotnik for school grades 9 to 12. All pupils took one of two special tests in addition to a regular attainment battery. For about 130 pupils from each grade the additional test was of the ability to identify correct and effective written expression. For about 350 pupils from each grade the test was of quantitative thinking and involved some interpretation of graphical and tabular material. Both sets of experimental test material comprised subsets from parent tests parallel to a test in the main battery. No net context effect of any size was evidenced by the language test material. However, for the quantitative questions the mean estimates from the item samples were consistently larger than for the whole test. Feldt and Forsyth, speculating on the difference, gave as possible explanations,

- a decrement in motivation with test length, the quantitative item sample was only a quarter the length of the full test compared with a half for the language item sample
- or, and possibly more likely, the greater mental demands of the quantitative test led in the longer test to clear experience of failure with negative motivational consequences that were avoided in the shorter test
- or that the time factor had inadvertently favoured the item samples (against this was the fact that noncompletion was less than 1% in the main battery).

If either motivational explanation were correct this would not necess-

arily mean that performance on item subsets had lower predictive validity than performance on these items in a full-length test: the reverse could even be argued. It would mean that item-sampling norm estimates would initially require some form of corroboration. For tailored testing it would mean that item calibration from a long test might be suspect. Once item standards had been validated against longitudinal criteria in the usual way then for selection and allocation the calibration difference would be of no consequence.

Apart from an overall context effect a carry-over influence from the difficulty level of the immediately preceding item has been claimed - especially an error-proneness following failure. Huck and Bowers (1972) reviewed such claims and investigated the possibility of bias in estimates of item difficulty from such a cause. Course examinations were prepared in a variety of orders for 120 and 160 psychology students. An analysis of variance procedure designed expressly for testing whether treatments (items in this case) have carry-over effects (Williams (1949)) was employed but did not detect such effects.

The research reviewed in this Section clearly allows the possibility that in some situations at least item characteristics are context-free. This is sufficient for the immediate purpose. Transferable item characteristics are best obtained from untimed administrations of short tests. Multiple-choice vocabulary (synonym) items are among those which have shown (for students) immunity to context. The raw data to be used in this thesis are an exact fit to the above prescription. It may be that in the psychometric theory that will evolve for tailored testing a place should be reserved for indices of context-reliability.

C. Background tailored testing research

1. Introduction.

In this section three kinds of research will be discussed. All are directly concerned with tailored testing but have in common that they are somewhat distant from the main line of this thesis. Sometimes the distance results from a difference in approach, sometimes it reflects the vastly greater computing power available today. The three kinds of research are,

- work using the sequential analysis procedures outlined in Section A. These are the procedures associated with Wald and were described above as leading to research of lesser relevance
- work before 1970 based on a fixed step up-and-down method that steers a testee through a pre-determined lattice or network of paths between items. Tests of this kind and period were usually referred to as branching or programmed tests. This nomenclature reflects a general influence or push from the then topical field of programmed learning and teaching machines
- work, possibly rich in ideas, but limited in its scope by the limitations of the technological facilities used or available. Such work (as reviewed here) was carried out before 1970.

The above classification simply defines what is being regarded as background research: it is not a division which can always be followed in the review below.

Before moving on to the earliest tailored testing research it will be useful to distinguish the following ways used to collect data.

Theoretical studies attempt within the limits of mathematical tractability to model a test situation. Mathematical functions which might show or have shown working approximations are used to explore tendencies, relationships and limits. The range of theoretical studies possible has been considerably extended by the availability of computers capable of executing solutions by numerical methods for the less tractable situations. Within the limits of their assumptions such methods are very powerful. Monte Carlo simulation studies generate test data from a theoretical base. This data will be a planned sample from the given area and will help explore a situation too complex or difficult to explore more exhaustively by theoretical means. Real-data simulations are based on data from encounters between real people and real questions. Such data is used as if it had occurred in a tailored test. In this way a sequence of individually selected items may be taken from a testee's test record with no regard to the original test order of items. This thesis makes extensive use of real-data simulation. Empirical studies are real-life tailored tests, presenting real people with real items and tailoring the choice of items to the individual person according to the procedure being investigated.

Empirical research with computer assistance is an expensive undertaking and usually follows only after preliminary research by one or more of the other methods. The order of presentation of the four methods is generally one of decreasing research accessibility. On the other hand attempts at pencil-and-paper implementations of tailored testing are not expensive and have been embarked on without preliminaries in a number of studies to be described.

The remainder of the section is a study-by-study review with interspersed summary views and comments. For these earlier studies it

is often the case that other details become as important as the results proper. These instances provide particular pegs for comments to hang on.

2. Tests using sequential analysis.

Let us take first for review research making use of Wald's sequential analysis. The earliest application to educational and psychological measurement was that of Cowden (1946). Perhaps unsurprisingly his was an empirical study with a class of statistics students. Grades for the course were assigned by a sequential procedure using a pool of 200 items from which subtests of 20 items were administered separately as conventional pencil-and-paper tests. Each subtest was marked before students went on to the next. Students only went on to a further subtest if - in the sequential analysis method - their performance so far had not classified them with sufficient confidence. He found three subtests were sufficient to classify a majority of students.

Moonan (1950) used a real-data simulation from responses to a 75-item achievement test. He investigated how well an item by item sequential analysis could approximate the pass/fail classification based on the whole test. On the average 40 items showed a good approximation.

Anastasi (1953) and Burgess (1955) reversed the roles of testee and item. They used sequential analysis to classify items for test suitability on the basis of a series of responses by different people.

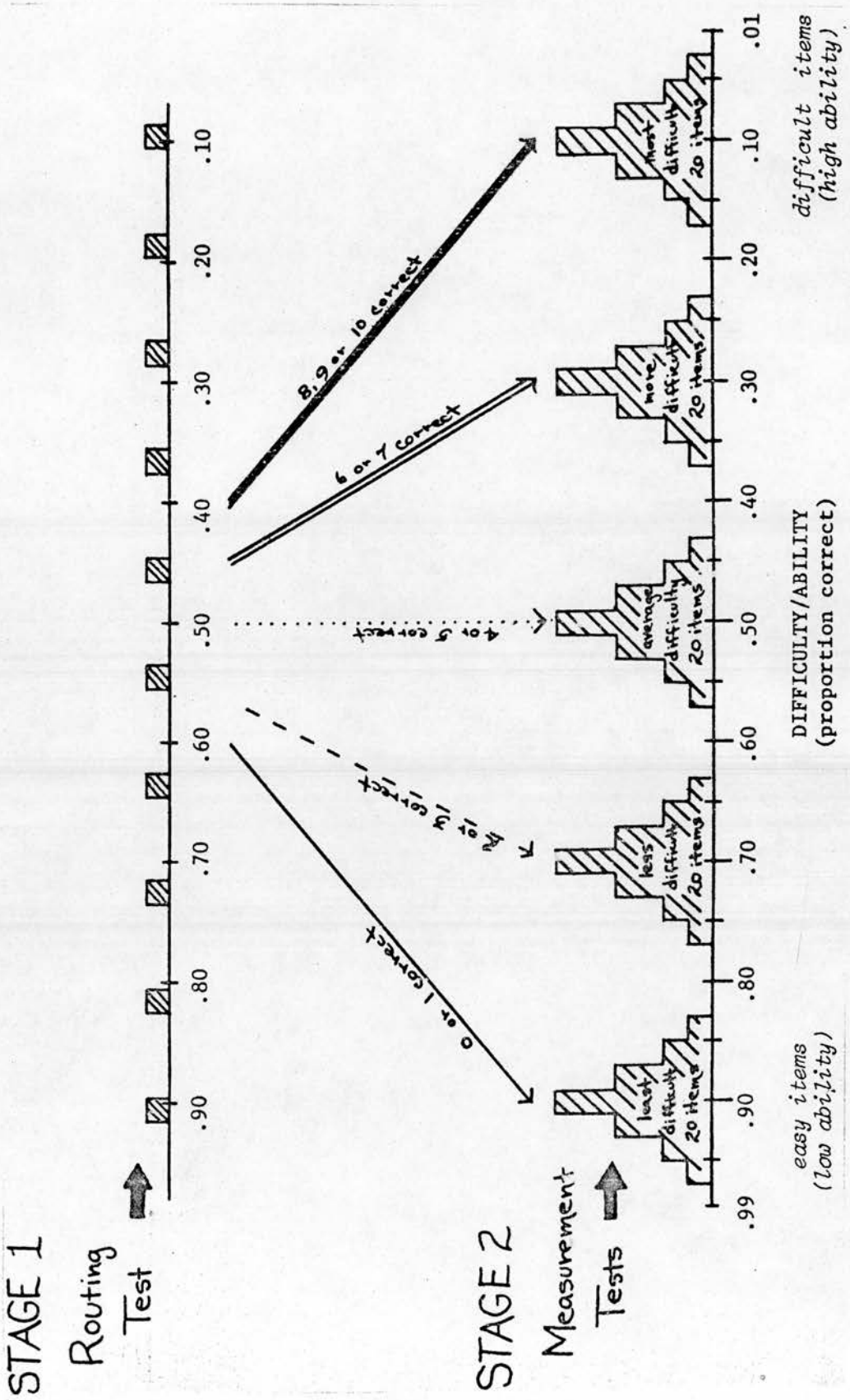
None of the four early researches above apparently offered a persuasive utility as the next tailored testing studies to use sequential analysis were not until 1968. So although the approach had been demonstrated in psychometric applications - and to some effect -

it was not perceived as useful.

A group of researchers (Cleary, Linn and Rock) experimented with variations of an elementary form of tailored testing. This form is two-stage testing. Here testing is in two parts. The first stage is common to all testees and is aptly termed a routing test as its function is to steer or allocate testees to the most appropriate of several tests making up the second stage of the procedure. The tests in the second part are relatively specialised, say by ability level, and are referred to as measurement tests. Figure 1 illustrates a two-stage procedure in which a 10-item routing test directs a testee to one of five 20-item measurement tests. Two-stage testing could be used with pencil-and-paper tests, especially if there were a little time between stages. A screening test governing admission to a full test battery could be viewed as a special application of a two-stage strategy, but more typically both routing and measurement tests are short by conventional test standards and all testees proceed to the second stage.

Cleary et al used sequential analysis for some of their routing tests and with some success. (Other forms of two-stage testing are discussed later.) Their technique was that of Armitage (1950) in which allocation to measurement test depended on the cumulative value of a probability ratio statistic. Real-data simulation from responses to items in scholastic tests taken by large samples of 11th-grade pupils and college students allowed an item-by-item consideration of performance on the subset of items selected to make a routing test. Of course, contrary to the requirement of Wald's approach, the items in the routing subset were not equivalent and this was recognised by the researchers; the items differed in both difficulty and discrimination so that responses could be regarded as random trials of a random

FIGURE 1 An example of a two-stage test (after Weiss, 1974).



variable only to an approximation. However, this approximate sequential analysis strategy was among the more successful of the routing possibilities explored. In the first study (Cleary et al, 1968 a) "sequential item sampling" was one of four routing methods tried. (Their other methods are referred to later.) Of these methods " the sequential method resulted in the fewest errors of classification and the highest overall correlation with total test score for both the original and the cross-validation samples", (p. 357). However, correlation with total test score was high for all four methods - ranging from 0·91 to 0·96 for the cross-validation sample - and generally only comparable with what the study also showed could be achieved through the use of shortened conventional tests using the best items.

The use of total test score in this study as a criterion for comparing alternative approaches is a device common in real-data simulation studies. It is the score on the conventional test that provides the basis for the simulation. As one estimate of the characteristic being assessed it is clearly appropriate to look at how well total score corresponds in turn with estimates by alternative means. Nonetheless, as a criterion, total score on a conventional test has limitations. Reproduction of conventional test estimates is not the prime purpose of tailored testing. Both conventional testing and tailored testing have the common aim of assessing psychological characteristics, and both no doubt can be expected to achieve this less than perfectly. Consequently while useful as a screening criterion a conventional test score is inappropriate for finer evaluations. By definition an improved method of assessment (or an equal but different method) will have a high but significantly

imperfect correlation with existing methods. The research of Feldt & Forsyth (1974) mentioned in Section B suggests, for example, that conventional and tailored testing might in some cases differ in their susceptibility to motivational influence.

In Cleary et al's study it should also be noted that correlations with total score carried a part/whole inflation (40 items out of 190). The simulated shortened conventional test also carried the same inflation so that comparability was not lost. However, an independent total score could well be used.

A follow-on study (Cleary et al, 1968 b) was restricted to sequential item sampling used to route testees to one of either three or four second-stage measurement tests. To achieve the same (inflated) correlation with total score (0·96) as that found for an average 37 items in the two-stage test required at least a 50-item conventional test.

Linn et al (1969), using the same real-data base, compared the same sequential item sampling strategy with other approaches against external criteria - in this case subsequent achievement test scores. The other approaches included branching as well as two-stage forms of tailored testing and these again are discussed further below. Against the external criteria all the tailored testing forms correlated more highly than conventional short tests made up to the same length from the best items. Of the several tailored testing forms those incorporating sequential item sampling were among the more successful.

Finally in this series of researches Linn et al (1972) used the same sequential testing procedure in a real-data simulation from college student examination response data. On this occasion the success

of sequential testing in classifying students into lower and upper groups was examined. Figure 2 illustrates their results for a mathematics examination. The increasing values for A in Figure 2 refer (not numerically) to decreasing levels of misclassification risk. The mathematics examination was 75 items in length. In this study the sequential testing took items in the same order as in the examination. Generally sequential testing required about half the items needed by conventional tests for the same number of correct classifications. For sequential testing it is average number of items required that is plotted; students away from the cutting point would generally need fewer than this average, those closer would need more.

Results in two other examination subjects were similar. These results agree closely with those of a theoretical study by Green (1970) which are illustrated in Figure 3 for a sequential test used to classify Ability Level as less or greater than a standard score of zero.

Working in a context better suited to the requirement of equivalent items Ferguson (1969, 1971 a&b) also employed a sequential analysis approach. The context was individually prescribed instruction and he was working on the assessment of proficiency in learning objectives. Especially at the elementary level, and perhaps especially in mathematics, it becomes possible to formulate item generation procedures (for example, to produce items calling for the addition of two 2-digit positive numbers less than 50). Where this is possible computer assistance can be used to generate further equivalent items to a built-in specification as they are required for testing. Figure 4 is the classic quality control chart as applied by Ferguson to proficiency testing. Ferguson (1971 a) describes an application of

FIGURE 2 A simulated comparison of sequential tests and short conventional tests. (Results from Linn, Rock & Cleary (1972): a mathematics examination is the basis here for assigning 2420 college students to lower and upper groups.)

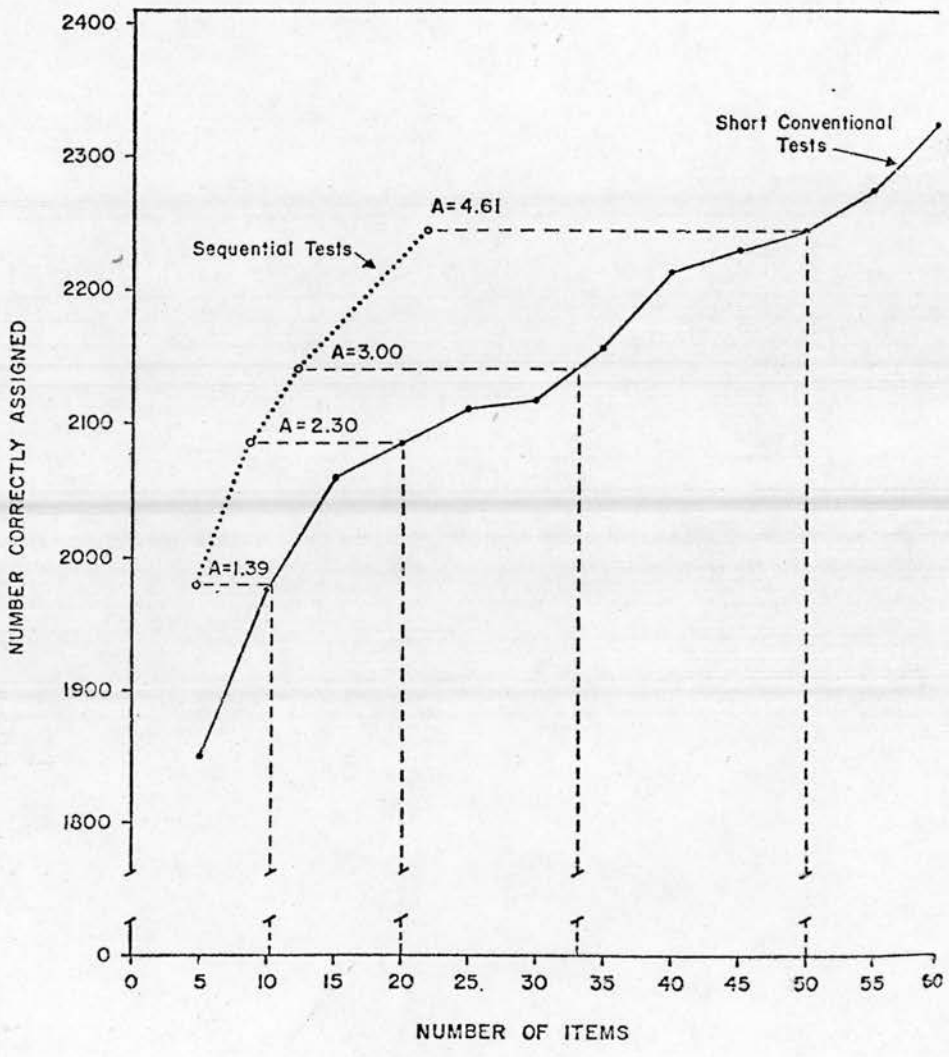


FIGURE 3 The number of items needed by a sequential test to match the operating characteristics of conventional tests n items long for the decision Ability above or below scale zero.

(Theoretical results from Green, 1970)

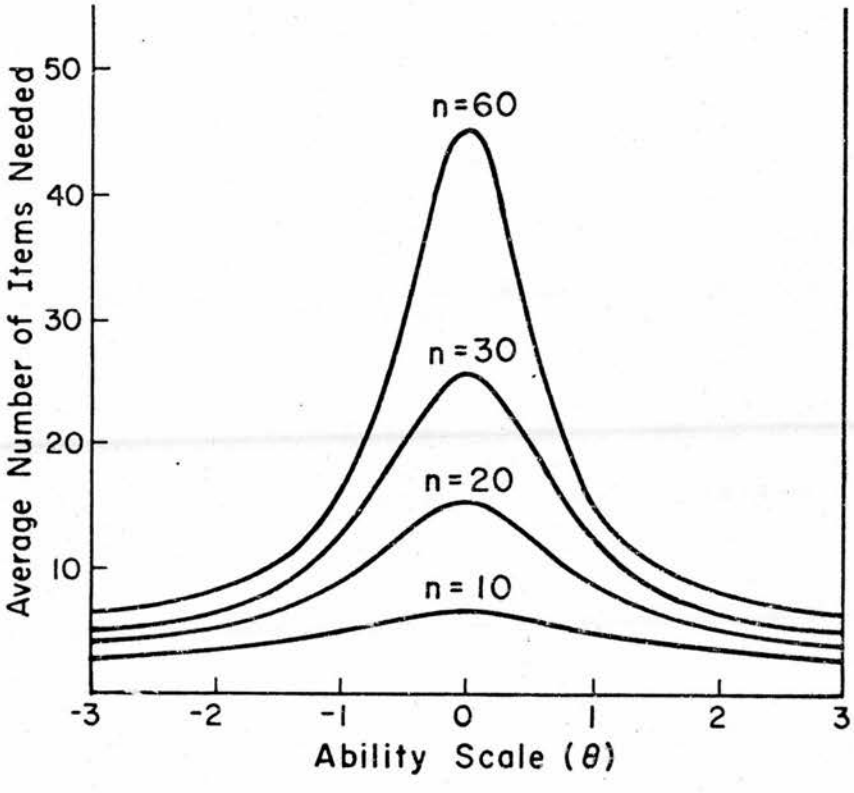
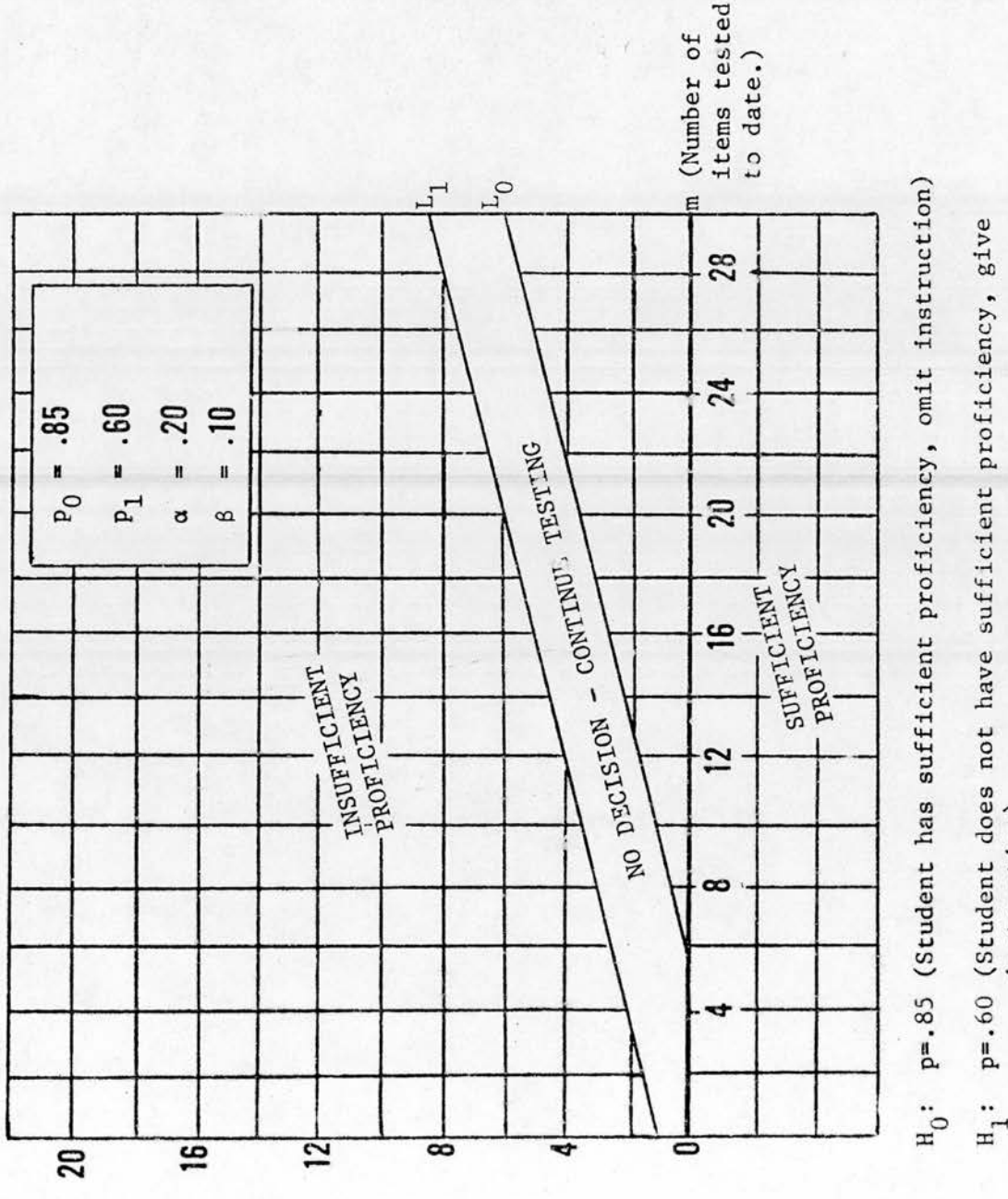


FIGURE 4

An application of sequential analysis to proficiency testing.

If p is the (unknown) proportion of all equivalent items that would be answered correctly by a pupil, then p_0 is the highest level for p for which a wrong-reject decision is judged serious, and p_1 is the lowest level for which a wrong-accept decision is judged serious. α and β are the risks accepted for these two decisions.

(after Ferguson, 1971 a)



H_0: p=.85 (Student has sufficient proficiency, omit instruction)

H_1: p=.60 (Student does not have sufficient proficiency, give instruction)

the sequential approach using item generators for testing various levels of addition-subtraction proficiency in an empirical study with pupils in grades 1 to 6. Questions were presented on a computer teletype terminal and responses made on a partially covered keyboard. No practical difficulties were reported. Branching rules for moving to the next objective were written so as to allow skipping up the objective hierarchy when high proficiency was established. Branching reduced the testing time required (although from an educational standpoint a more important finding was that more items were generally found necessary for proficiency decisions than the conventional test procedures had allowed). Assessments from the sequential procedure were judged as valid and reliable as those from conventional tests.

Sequential analysis procedures have shown benefits in applications calling for assessments to divide people into two (and possibly three and four) subgroups on either side of a pre-determined cutting level. This situation is likely to arise in educational or training programmes in relation to mastery of units of instruction: on the other hand the method would not cope comfortably with the provision of diagnostic information in the case of non-mastery. A possible use in selection would be for the initial screening of job applicants where minimum qualifications on critical abilities and attainments could be tested in this way. However, sequential analysis is not an appropriate method for helping the general allocation of personnel, although its explicit formulation of misclassification risks is regarded as a desirable feature.

3. Two-stage tests.

Now we will return to look more closely at the two-stage testing

procedures introduced above.

In a large scale empirical study Angoff & Huddleston (1958) compared two-stage college entrance tests in verbal and mathematical aptitude to conventional tests. In both subjects a routing test "directed" pupils to one of two measurement tests. In fact, using a sample of 6,000 pupils all the combinations of measurement and routing test were used so that a subsample necessarily took the appropriate measurement tests as if routed. The measurement tests were more reliable than the conventional tests, and showed slightly higher predictive validity against grade point average. The routing procedure made some 20% of routing misclassifications. The technical superiority of the two-stage procedures was not considered sufficient to offset the administrative difficulties that would arise.

In research already referred to for its use of sequential item sampling for a routing test Cleary et al (1968 a) also experimented with three other routing tests. All routing was to one of four 20-item measurement tests. This study was a real-data simulation using responses of several thousand 11th grade pupils to 190 multiple-choice verbal items. The three routing methods were:-

1. Double routing: A 10-item initial test was composed of items of about 50% difficulty level. Scores on this test were used to divide the sample into two approximately equal groups who went on to two separate 10-item tests similarly constructed in relation to their own groups. A further split then directed testees to the four measurement tests.
2. Broad range routing: A 20-item routing test having a rectangular distribution of item difficulties (as illustrated in Figure 1)

divided the sample into approximate quarters based on the 20-item score.

3. Group-discrimination routing: The total sample was divided into approximate quarters on the 190-item total score. Item difficulties were then evaluated within each of the four groups. The 20 items with the largest difficulty range between top and bottom quarters were then selected for the routing test. Allocation to measurement test was on the 20-item score.

The last approach is interesting in that it explicitly recognises in a small way that tailored testing may require other item parameters than are appropriate for conventional test construction: in this case item difficulty by a coarse ability grading was used rather than overall group difficulty (an approach not unknown in conventional work but less common at least). That the different approaches give different results is shown by the following details. The 20 items selected for the group-discrimination and broad range routing tests had only six items in common. The sequential item sampling routing test (described earlier), made up of the 23 items having the highest point-biserial correlations with total test score, had only 10 items in common with the group-discrimination test.

For classifying the total sample into quarters compared with the "true" 190-item classification, group-discrimination routing (29% misclassifications) was clearly superior to broad range (39%) or double (41%) routing. Sequential routing (27%) did slightly better. In terms of reproducibility of the 190-item score similar relativities obtained between correlations of this score and the four two-stage approaches. However, only the sequential approach was as effective as a 40-item conventional test.

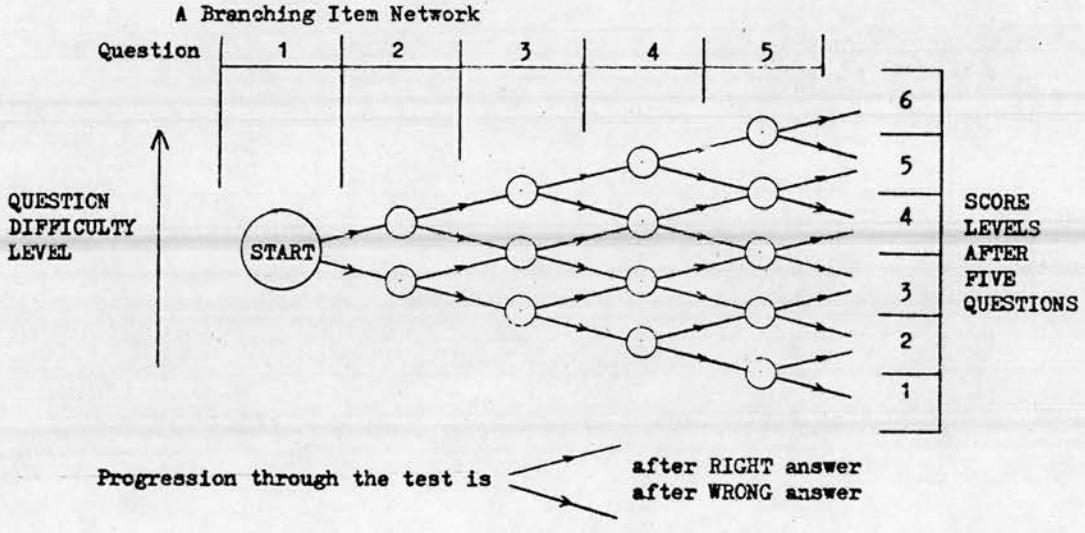
In Linn et al's (1969) follow-on research using external criteria (described earlier) the group-discrimination approach was superior to the other approaches in allowing prediction of these criteria, and much superior to a conventional test of the same length. A conventional test three times the length would be needed to give comparable results.

The two-stage testing research reviewed is generally encouraging. One would like to see Linn et al's (1969) favourable results confirmed in empirical studies before accepting the absolute size of the advantage. It may be significant that the most favourable result was achieved by the group-discrimination approach which looked a little beyond conventional item statistics.

4. Branching tests.

Next are the earlier fixed-step up-and-down sequential procedures. These procedures form an evolutionary line which continues through into Section D of this review. The archetypal procedure is based on a branching network of pathways through a fixed lattice of questions. An example is illustrated in Figure 5, but there are many variations. All testees begin with the same START question, usually of middle difficulty, and move through the network along routes which depend on their performance on successive items. Any testee will be steered through only five of the fifteen questions in the network. Referring to Figure 5, the more able testee will tend to get his initial questions right but after branching upwards to questions of greater difficulty he will find a better match. "Fixed-step" refers to the constant difference between neighbouring difficulty levels; "up-and-down" refers to the method of steering. In a more extensive network

FIGURE 5 An example of a branching test.



than the 5-stage plan of Figure 5 it will only be extreme testees who by the end of their test have not been encountering items approximately matched to their ability. In the later stages of such tests the answers of most testees can be expected to show a rough balance between wrong and right. In this way the test taken is tailored to suit the individual testee. Such tests are variously referred to by later researchers as branching, programmed, or pyramidal tests: an earlier term was sequential item tests, but to avoid confusion with sequential analysis methods this term is not used below.

Krathwohl & Huyser (1956) were the first to employ a branching test of this kind. Interestingly they had been looking at a sequential analysis approach (which we have already seen achieved earlier adoption) for switching testees from one block of questions to another. However they came round to the automatic-routing design of the branching test. Their thinking, of course, had a pencil-and-paper context in mind where sequential analysis sets administrative problems. First they used a real-data simulation from a 60-item college-level ability test. The test had 5-option multiple-choice items and Krathwohl & Huyser distinguished not only right and wrong answers but also better and poorer wrong answers. Their branching test had three exit paths from each item rather than the two of Figure 5. Because guessing in the conventional test data base seemed to be raising branching test scores unduly a new design was tried for further simulations. This design had two items at each node in the branching network - a block design in bioassay terms. Again there were three exits from each node, depending this time on whether two, one, or none of the items there were correctly answered. Correlations of about 0·77 with total score were obtained by a three-stage branching test of this kind which

considered only six of a student's 60 responses.

Subsequently Krathwohl & Huyser tried an empirical pencil-and-paper implementation of their scheme, the most important outcome of this trial being that they ran into considerable practical difficulties in test administration.

The United States Army took up research on branching tests primarily with the aim of finding shorter tests. This is reported in a number of studies from 1960 onwards. Bayroff et al (1960) and Seeley et al (1962) constructed four 6-stage branching tests to a modified Krathwohl and Huyser design. Then in an empirical study they tried out pencil-and-paper implementations of two tests - verbal and arithmetic reasoning. The branching tests were administered to 327 enlisted men. Despite finding that the tests were too easy (no suggestion of motivational causes was made for the high scoring) correlations of 0·68 and 0·74 respectively were found for the 6-item branching tests with independent parallel 50 and 40 item conventional tests. On the other hand Seeley et al (1962) also concluded (p. 7), "..... it became apparent that the SIT [the branching test] possessed some characteristics not entirely advantageous in terms of intended Army use." They went on to detail these as follows:-

1. The branching test was more costly and time consuming to construct.
2. Administration of the 6-item branching tests was lengthy. For the two tests 10 to 15 minutes of initial instruction were required as well as the 15 minutes allowed for test completion.
3. Scoring presented problems as a testee's self-routing through the branching test had to be checked.
4. The instructions for the branching test were not understood by substantial proportions of men. Overall 9% of the verbal and

21% of the arithmetic reasoning test records were not scorable (note that the arithmetic reasoning test was attempted second within the single 15 minute time limit).

As might be expected the proportion of not-scorable records was related to performance on the Armed Forces Qualification Test. Men in Mental Category IV (10th - 30th percentile) had the highest proportion of not-scorable records.

The researchers suggested that further experimentation with branching tests in this form was not worthwhile, but that the basic concept may have considerable utility for presentation via a testing machine. Bayroff (1964) reported a feasibility study for a programmed testing machine but this was not then built although reported as within the state of the art. (However, it will be seen in Section D that Bayroff et al (1974) do develop a programmed testing system.)

Leaving the US Army studies temporarily, Patterson (1962) had explored widely at a more abstract level. He used a computer-assisted Monte Carlo simulation applied to 6-item conventional and branching tests. The limitation to such short tests was imposed by his computing facilities. Within his branching test he placed the most discriminating items first within their difficulty level. He departed from a fixed difficulty step between items by allowing higher item discriminations to call for a larger step in difficulty level in the choice of the subsequent item. He also studied the influence of item discrimination, and of the shape of the ability distribution assumed. He found that his branching method gave more precise ability estimates for more extreme levels of ability, but that overall there was little to choose in precision against the conventional test. The branching

test results reflected non-normal ability distributions more sensitively. Errors in estimating the item statistics were found not to be critical.

For the US Army Waters (1964) carried out a theoretical study comparing 5-item branching and conventional tests. She assumed a normal distribution of underlying ability and normal ogive item characteristic curves. For both open-ended and multiple-choice questions she showed that branching test scores correlated more highly with underlying ability than the best of various conventional tests. The difference was small - of the order of 0·03 (open-ended) and 0·01 (multiple-choice) on coefficients around 0·8. Whether this advantage would increase with more extended branching tests and at what test length (if any) such advantage would dissipate were unanswered questions.

Bayroff and Seeley (1967) in a further empirical study of branching tests - but now with computer assistance - administered 8-stage branching tests of verbal and arithmetic reasoning abilities to 102 enlisted men. (The most able testees also went on to a 9th item.) This was possibly the first example of a computer-assisted tailored test. Test items were presented to individual men using on-line teletype computer terminals. Responses were made on the keyboard - the items were multiple-choice so that only option identification was called for. Correlations of branching test scores with independent 50-item verbal and 40-item arithmetic reasoning conventional tests were 0·78 and 0·74. Short conventional tests would need to be twice the length of the 8-item branching test to achieve comparable results.

In the British Army McGill (1968) reports the construction of a 10-stage branching test under the supervision of K.D. Duncan. The test was constructed from the multiple-choice items of a predominantly

mechanical aptitude conventional test. The most and least able testees were provided with further stages beyond the tenth. A real-data simulation from recruit response data showed close agreement with corresponding 60-item parent test scores. Further work then followed to produce a manageable pencil-and-paper format for empirical study. A technique that seemed to offer promising simplicity was one using an answer sheet over an embossed card so that embossed numbers would appear on shading a chosen answer space with a soft pencil (following Duncan (1964)). The number which appeared directed the testee to his next question. Small scale partial trials were reported to be successful.

Hansen (1968) carried out two empirical studies of branching tests presented by online teletype. His subjects were university freshmen taking a physics course examination. In his first study 56 freshmen took five topic-centred 3- and 4- stage branching tests, 17 items were attempted in all. Hansen also explored a variety of scoring methods. So far in this review only the straightforward scoring scheme illustrated in Figure 5 has been introduced for branching tests. There are other possibilities and these are discussed in Section D in relation to more recent work. Generally the scoring methods intercorrelate highly - the four methods used by Hansen had intercorrelations from 0·84 to 0·94. The validity of the four scoring methods for predicting final course grade ranged from 0·38 to 0·49 - all higher values than achieved by a 20-item conventional classroom test also taken by all students. A second study, also small (30 freshmen), is of interest because after the teletype test sessions the students completed an attitudinal scale about computer-based testing as they had experienced it. Generally their ratings were favourable.

Guessing was reported as happening very seldom. Disappointingly for putative motivational benefit students reported that they were relatively unaware of the efforts to individualise the test material.

In the real-data simulation study by Linn et al (1969) referred to previously two branching tests were included in the methods tried. One branching test was a normal 10-stage network, but with a weighted scoring system in which more difficult questions had higher scoring weights. The second test was to a block design. A block of five verbal items occupied each node in a 5-stage network. Testees thus attempted 25 items. The five items at a node were closely similar in difficulty. Branching from a node depended on whether two-or-less or three-or-more of the five items were answered correctly. Again a weighted scoring system was used. For predicting an external test criterion it was found that conventional tests would need to be 1·65 and 1·76 times as long as the two branching forms respectively.

Finally, in the pre-1970 branching test studies, Wood (1969) made up three tests of four, five and six stages on CSE mathematics topics. These were administered in an empirical study to 91 CSE candidates. The method of presentation was an improvised pencil-and-paper technique using self-adhesive labels. Wood experienced about 5% of spoiled papers. The correlations between the summed branching test scores (15 items in all) and subsequent CSE grade was 0·51 - compared with an almost identical value, 0·52, found for a short conventional test composed of the 15 best items.

So far the research on branching tests has shown persistent glimpses of possible benefits among a variety of cautionary results. Branching tests have in some instances,. and often by small margins,

nudged in front of equal length conventional tests in their relationship to underlying ability, their validity, their precision of estimate for non-average levels of ability, and in their reproduction of independent conventional test scores. Such encouragement was at least sufficient to sustain the converted.

A number of empirical studies were reported, all on smallish samples for obvious reasons. The later pencil-and-paper formats go some way towards relieving the despair of the first proponents and appear usable in some applications. The online use of computer terminals resolves the administrative problems most satisfactorily. All the terminals used have been teletypes, but their relative slowness and noise have not attracted any adverse comment.

Research so far has been concerned mostly with short branching tests, 10 items or less. This restriction is partly because it is early research on a topic but is also partly deliberate as a search for short test forms. A short-test viewpoint may emerge as a rather blinkered perception of the possibilities. Tailored testing can be more than an abbreviated substitute.

A heavy reliance on correlational methods of evaluating tailored testing approaches (and evident in this Section) was criticised earlier from the stand-point of the mere reproducibility of other estimates being an insufficient criterion. Correlational evaluations have also been criticised (Lord (1970 b), Wood (1969)) on the grounds that

- the correlation coefficient is a group statistic while for an individualised method of testing the focus should be on individual accuracy
- the value of a correlation coefficient is dependent upon

the distribution of the characteristic in the particular group.

Consequently although it is entirely appropriate to look for predictive validity in a tailored test estimate the force of the criticism is that this necessitates looking beyond a group correlation. The matter of evaluation is important and is taken up again in Section D.

D. Recent tailored testing research

The research reviewed in this section is that which, with a few exceptions, has been reported from 1970 onwards. As compared with the work in Section C the more recent research is characterised by greater sophistication of theory and equipment, and by growing coherence: there is also more of it - there having been more work reported in this period than in all the years before. This Section draws the research together under the following headings:-

1. General - this part introduces a number of concepts and approaches which are generally helpful and are used thereafter.

Further research on procedures already met

2. Short tests
3. Two-stage testing
4. Branching tests

New procedures

5. Flexilevel tests
6. Item-finding procedures
7. Stradaptive and broad range approaches

The testing strategies to be reviewed in parts 6 and 7 are of greatest relevance to the present thesis.

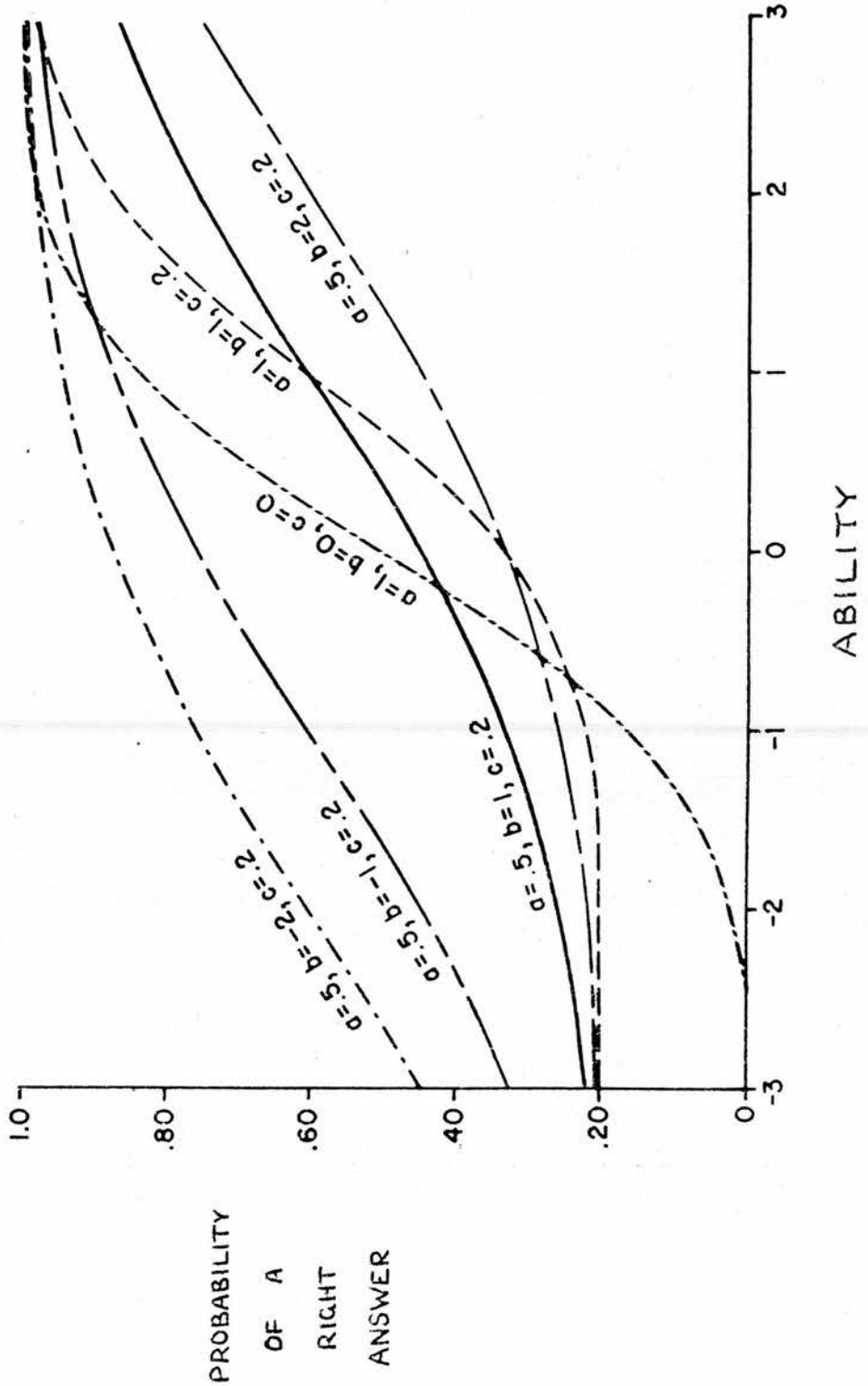
1. General

A number of researchers base their approach on latent trait mental test theory, or item characteristic curve theory as it is perhaps more descriptively also known. Figure 6 illustrates a number of item characteristic curves. Each curve represents the probability of success on a particular test question in relation to ability level. The basic theory assumes the curves to be normal ogive or alternatively logistic functions and is given in Lord (1952), Birnbaum (1968), and Lord & Novick (1968). In terms of numerical outcomes the choice between the alternative functions is of little consequence. The fit of the models to test item data has been evaluated in a number (but not a large number) of studies (for example, Lord, (1970 a)) with positive results.

Each item characteristic curve is specified by three parameters, a, b, and c together with the function assumed. Parameter c is the probability of chance success on a question: in multiple-choice questions c is often taken to be the reciprocal of the number of options although this is a questionable assumption; more empirically c may be estimated from the asymptote approached by the curve as ability decreases. The parameter c represents a considerable theoretical complication and its estimation for real-data requires large scale computing facilities and even so difficulties remain. When c is taken to be zero - this is realistically so for open-ended questions - parameter b can be simply defined as the ability level for which the probability of success is 0·5. b can be regarded as an index of item difficulty. More generally, when there is an appreciable probability of chance success, b is the ability level corresponding to the point of inflexion on the item characteristic curve. The remaining parameter a, can be taken to represent the discriminating power of the

FIGURE 6 Examples of item characteristic curves.

(after Lord, 1969)



item. Graphically the more discriminating items have steeper item characteristic curves. Parameter a is related to the slope of the curve at the point of inflexion. (For the normal ogive a is the reciprocal of the standard deviation.) Figure 7, from Urry (1971 b), presents values of a and b in relation to the conventional item statistics of proportion passing and point-biserial correlation with total test score: the figure is for c set at 0·2 (as may apply for a 5-option multiple-choice item) and this accounts for the asymmetry.

Lord (1974 a) helpfully reviews the relationship between tailored testing and item characteristic curve theory. Generally the theory offers a useful framework for real-data or Monte Carlo simulations and several studies of this kind are described below.

A further concept of general utility in this Section has to do with the evaluation of tailored testing procedure. It has been seen already that evaluation may be furthered by correlations with conventional measures, and, in the case of theoretical or Monte Carlo studies, correlations with underlying ability and precision of estimate.

An additional form of evaluation is by the use of information functions, and in particular a function recommended by Birnbaum (1968) and Lord (1952). Referring to Figure 8 for illustration we are concerned there with the ability of the measuring scale to distinguish the two ability levels A1 and A2. The Figure shows for these levels the distribution of measurement errors around the expected values X1 and X2. The success of the scale in distinguishing A1 from A2 is clearly dependent

i. on the rate of change of X with A; that is the slope of the line P1P2

& ii. inversely on the dispersion of the error distributions.

FIGURE 7

The relationship between item characteristic curve parameters a and b and conventional item statistics when the probability of chance success is 0·2. (after Urry, 1971 b)

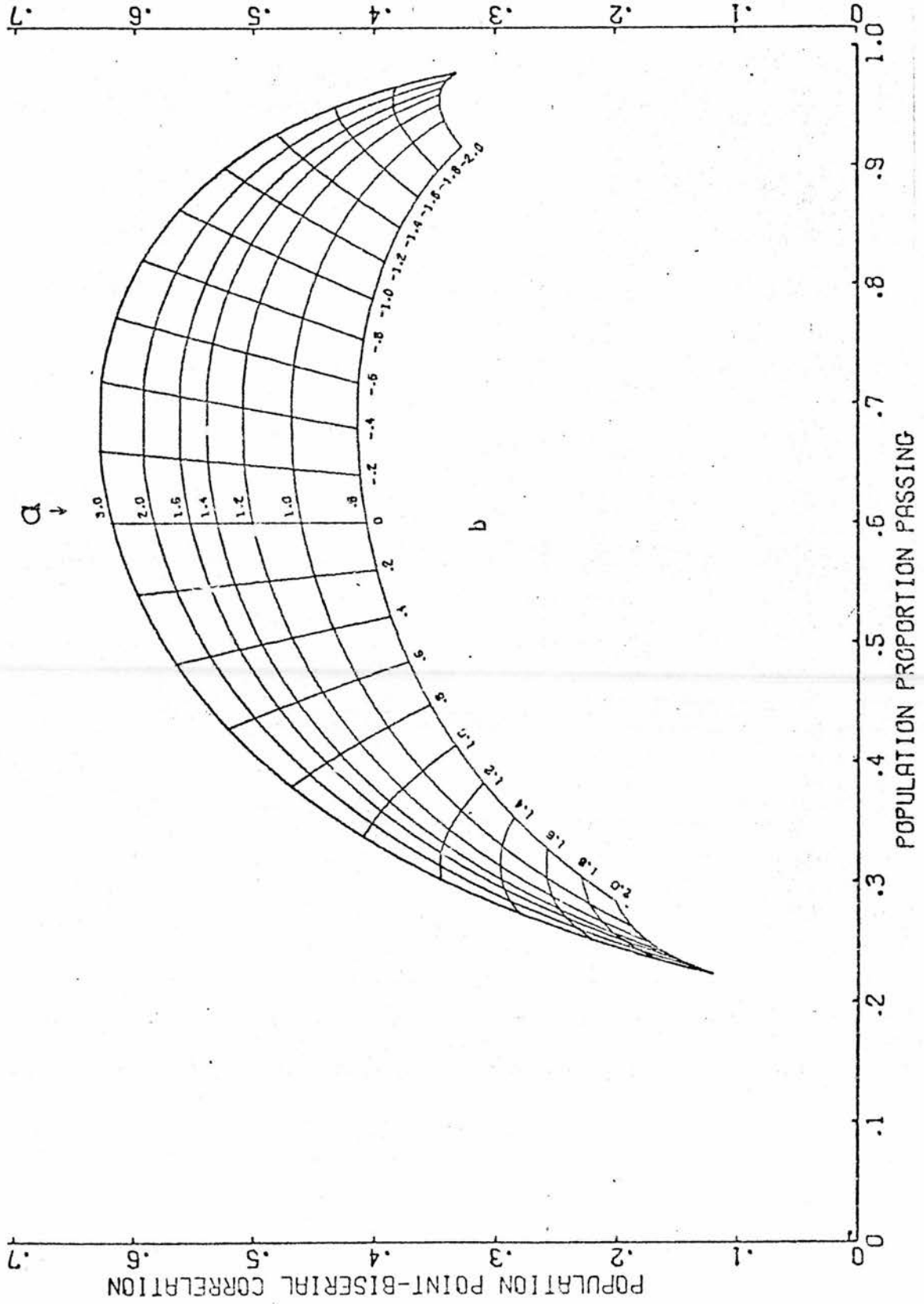
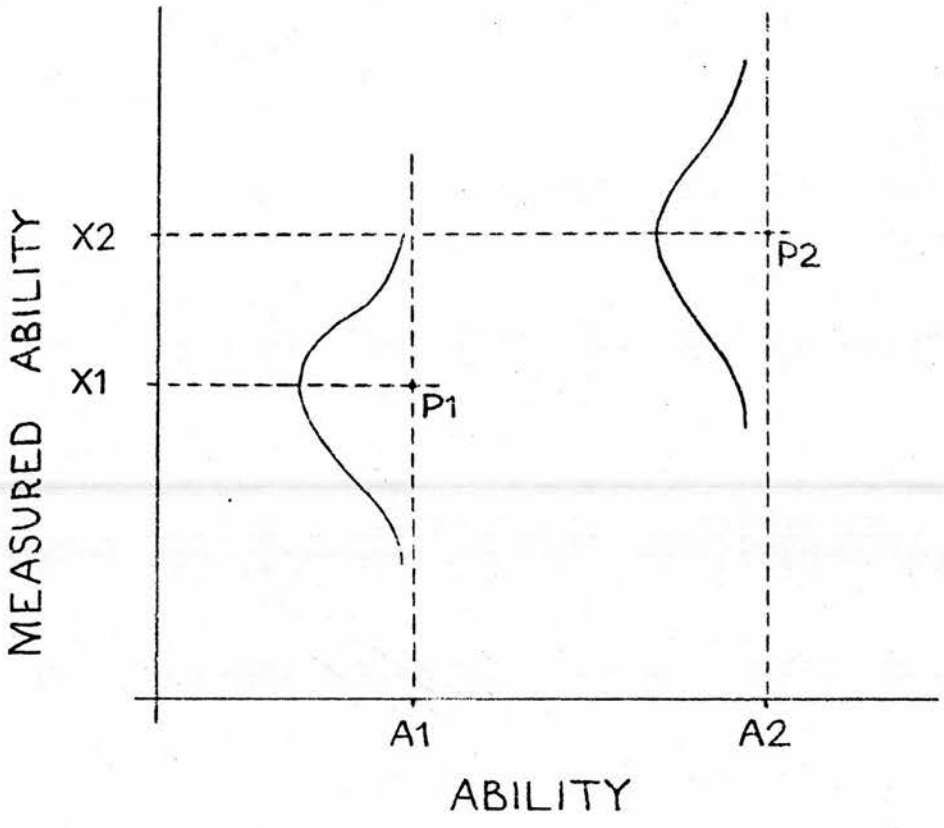


FIGURE 8 Estimating ability from a measurement scale.



The information function of interest is defined as the square of the ratio i/ii. Two other related interpretations of this information function have also been made. An increase in the information function achieved by a modified test design is the equivalent for a conventional test of a proportionate increase in test length. Secondly the information function, or more precisely its square root, is inversely proportional to the confidence interval for estimating ability level from test score. Where an information function is subsequently referred to it is this function.

Generally the information function is most usefully employed in comparing two tests by looking at the ratio of their information functions for different levels of ability. This ratio is termed the relative efficiency of the two methods and has the advantage, in this ratio form, of being invariant in relation to the idiosyncracies of the ability scale incorporated in the information functions. It follows that an absolute interpretation of an information function may be misleading. This danger is underlined by Lord (1975 a) who points to certain deficiences in the ability scale normally used in item characteristic curve theory.

In these general preparatory remarks the writer would also like to point to valuable reviews by Weiss & Betz (1973 a) and Wood (1973). Their preferred terms for tailored testing are adaptive testing and response-contingent testing. Weiss (1974) also presents a useful comparative commentary on the various approaches tried for tailored testing.

2. Short tests

The few studies reviewed here continue the predominantly military

concern of producing shorter tests. This abbreviation may be attempted by any means; individualised testing procedures are but one line of attack.

Bryson (1971 and 1972) looked at four methods of producing 5- or 6- item tests. Initially she used real-data simulation based on a response bank from 10,000 men in recruit training at a Naval Training Center. Responses to two tests were used, the Navy General Classification Test and the Navy Mechanical Aptitude Test. One of the four methods was a shrinking-step individualised branching procedure referred to as BRANCH. In BRANCH the question with the highest internal validity is first used to split the total group. Internal validities for the remaining questions are then recomputed separately for the two groups. For each group the question with the highest validity for that group is then used to make a further split. Thus the procedure routes testees through a series of forks so that after five questions there are 32 exit points.

Thus a major characteristic of BRANCH is that question selection is based on a question's local (rather than total group) characteristics for a sub-group of a narrower range of ability. This seems vital to tailored testing. Procedures based on total group item statistics only make sense in so far as these statistics offer approximations to the performance of the items for people of more homogeneous ability. Essentially tailored testing treats people differentially in relation to ability. It will be preferable to avoid the approximation from total group statistics (and the assumptions inherent therein) and to work directly with item indices related to ability levels.

A critical disadvantage of BRANCH is that it offers no recovery

route after an incorrect forking decision.

Compared with the other methods BRANCH was most successful in reproducing total test score. For the general classification test (the more internally consistent of the two tests used) correlations with total score for the four short test methods ranged from 0·86 to 0·94 for 5-item tests. For the mechanical aptitude test the range was 0·69 to 0·82.

Bryson (1971) went on to give empirical trials to the four short test methods. In these trials BRANCH tests were administered by online VDU terminal to 263 recruits. Each question was given with a separate (55 second) time limit. Under these conditions BRANCH was no better than the best of the other methods in reproducing total score. Bryson points to the original choice of BRANCH questions being based on item characteristics which for later items would be influenced by time pressures not present in the BRANCH VDU presentation. This is a likely factor and emphasises the importance of realism in any response base used for simulation.

Outside the context of short tests correlation with total test score is only a start in the evaluation of alternative procedures.

3. Two-stage testing

Two-stage tests are a marginal form of individualised procedures. Generally a two-stage test will offer perhaps only three to six alternative diets of questions. The saving virtue of a two-stage test may be that because pencil-and-paper implementation is possible it does offer a realisable prospect of large scale testing. Whereas large employers, such as the Army, maintain continuous recruitment so that online testing can be achieved with a relatively small

number of computer terminals, other settings - notably educational examinations - may demand the capacity to test large numbers of people simultaneously.

At a theoretical level Lord (1971 c) used item characteristic curve theory to investigate nearly 200 two-stage designs. He assumed a normal ogive characteristic curve and also equal discriminating power (constant value of parameter a) for his items. Largely he worked with an overall limit of 60 items for routing and measurement tests combined. He considered both no-guessing (c=0) and with guessing (c=0·2) conditions.

His basis for comparison was a 60-item peaked conventional test. By peaked he means a test in which all items are of identical difficulty. A test peaked at ability level A would be such that the probability of someone of ability A answering any one question correctly would be 0·5 (excluding chance success). Hence a peaked test differs from most conventional tests in regular use: such tests albeit geared to specified populations typically have a spread of item difficulty. The routing and measurement tests of his two-stage designs were also taken to be peaked at appropriate ability levels. A "best" up-and-down branching test of equal length provided another basis for comparison.

In scoring his two-stage test designs he used a maximum likelihood estimator. That is, assuming normal ogive regressions of item score on ability he determined (by large, fast computer) the ability for which the observed set of item responses was most likely. The information function already described was used to evaluate his results.

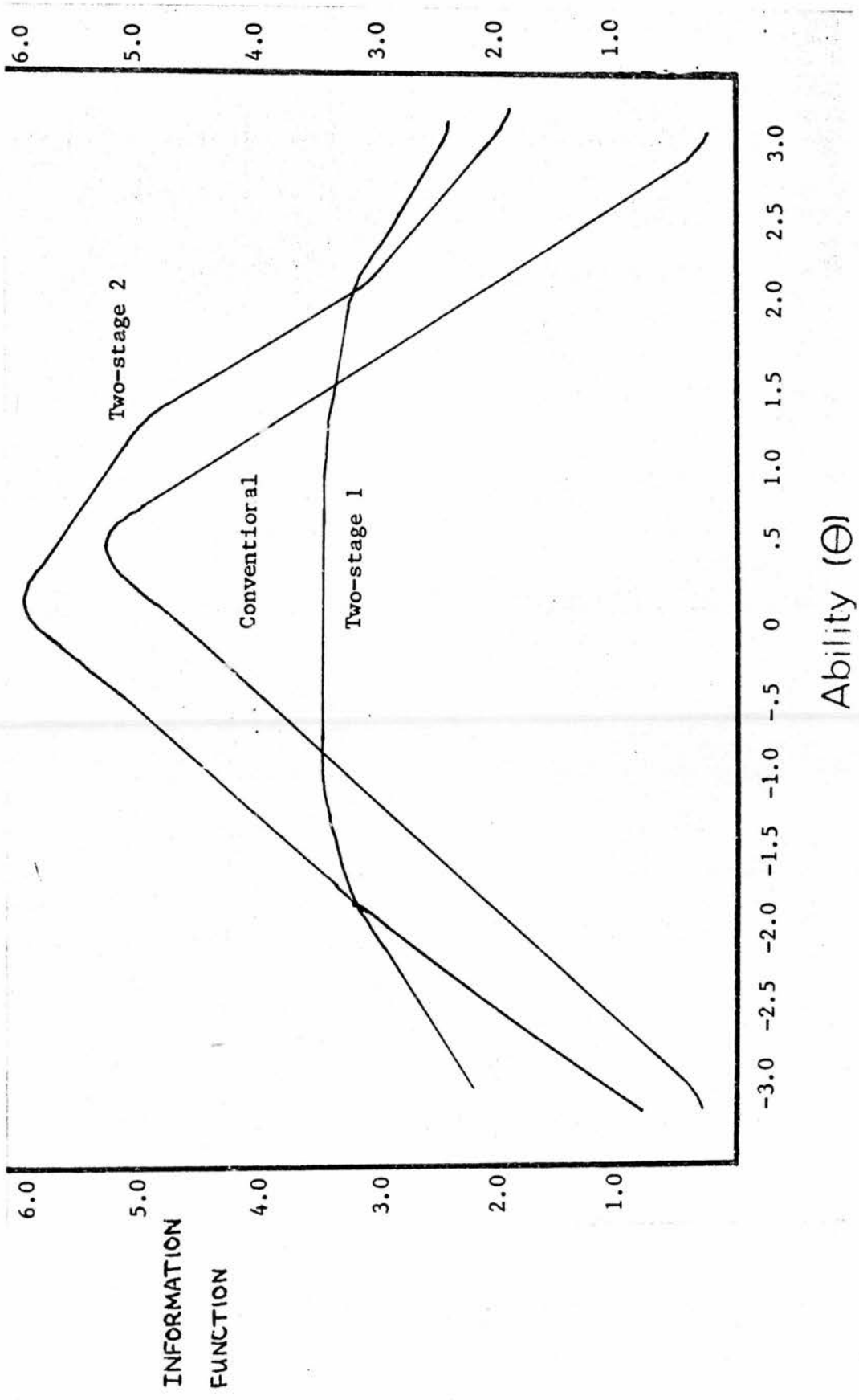
His findings indicate that with no possibility of chance success (c=0) the best two-stage procedures are as effective as the best up-and-down procedures. However, with c=0·2 no two-stage procedure was quite as effective as the up-and-down test. In both cases the peaked test was better at and around the ability level at which peaked but substantially poorer elsewhere.

Betz & Weiss (1974) carried out a Monte Carlo study following Lord's in a number of ways but keeping to a 40-item limit and using the item characteristics of an available item pool for their simulation. Hence their conventional 40-item test was not peaked in the narrow sense. Figure 9 summarises their results in terms of the information function over the range of ability. The authors point out that the "Two-stage 2" test used items with slightly better values of parameter a (discrimination) than the conventional test. Two-stage 2 is superior to their conventional test over the ability range. Two-stage 1 is superior at the extremes of ability.

In an earlier study using the Two-stage 1 design Betz & Weiss (1973) had carried out the first computer-administered empirical study of two-stage testing. 214 psychology students were tested using an online VDU terminal. Difficulties had been encountered with both the measurement tests and the cutting scores on the routing test that determined allocation to measurement test. These difficulties can be attributed to their use of total group item statistics. Two-stage 2 resulted from modifications to this first design.

From the early '70s a research group led by D.J. Weiss has worked on adaptive testing - to use their term - at the University of Minnesota and further references will be made to their work. In

FIGURE 9 A comparison of simulated two-stage and pseudo-peaked conventional tests. (Results from Betz & Weiss, 1974)



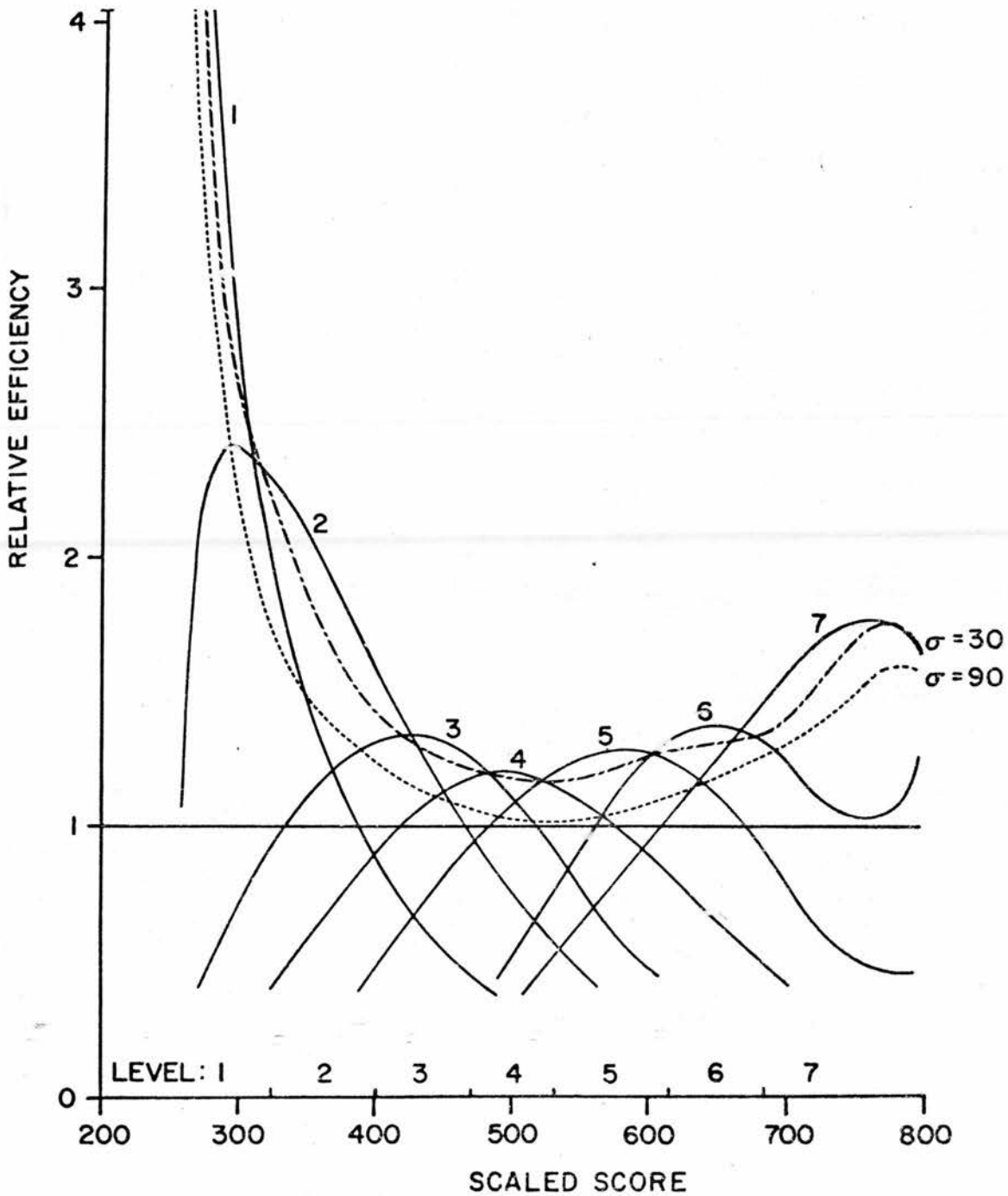
particular the group has started on a programme of empirical trials of various approaches to tailored testing using VDU online terminals. Empirical work in tailored testing remains rare and the experience of the Minnesota group, albeit confined to psychology students, often provides pioneering information on the topics covered.

In contrast with empirical difficulties Lord (1974 b) in an intriguing paper continues to tempt the applied researcher with attractive theoretical results. In so doing he demonstrates the power of item characteristic curve theory where its assumptions can be realised. In this study he is looking principally at the nature of the measurement tests in a two-stage procedure. He now refers to the concept as a multilevel test. For his multilevel test he refers explicitly to the desirability of item overlap between adjacent levels: the reason given is that of item economy, but it has seemed to the writer that in the interests of the individual who might be misrouted or misallocated to level that such overlap was very desirable to avoid a patchy kind of measurement superiority sprinkled with individual failure.

A College Entrance Examination Board SAT Mathematics paper was used to illustrate the approach. Figure 10 shows the relative efficiency of the seven individual level tests compared with the full length test. (The relative efficiency is the ratio of information functions described earlier in this Section.) Each individual level test is only two-thirds the length of the full test. The horizontal line at a relative efficiency of 1 is for the full test. The solid lines plot the curves for the seven individual tests: each such curve has a relative efficiency above 1 for its portion of the score range. The dashed lines give the overall efficiency of the multilevel test

FIGURE 10 Theoretical results on an SAT Mathematics paper comparing the relative efficiency of a two-stage multilevel approach and the conventional test. The dashed curves are the overall efficiency of the seven (solid line) local tests for two levels of standard error (σ).

(from Lord, 1974 b)



for two values of standard error of measurement in the routing test. A standard error of about 75 scaled score points would be achieved by a 12-item test. The upper curve for a standard error of 30 would not be practically attainable. Initial misallocations by one level is seen to be of little consequence but an error of two levels could in some cases lead to substantial relative inefficiency.

4. Branching tests

The basic concepts of branching tests have already been introduced and an illustrative schema was given at Figure 5. Probably more work has been done on such schemes than on any other individualised approach. This approach has the critical disadvantage of requiring a very prescribed item pool. Items are needed to fit the nodal points where routes meet and diverge. A branching test cannot be used until every node has an item of approximate fit. Neither are the item requirements negligible: $n(n+1)/2$ items are required for an n-item branching test - 120 items for a 15-item test. The test constructor would be considerably dismayed at the thought of how many items would need to be written to obtain the 120 to match the specification. Additionally the item specifications are made in terms of conventional total group statistics which at best can only be an approximate indication of performance for the relatively narrow ability band of testees encountering any one item.

Several theoretical studies have been carried out based on item characteristic curve theory. Often for simplicity fixed values are assumed for discriminating power and probability of chance success (parameters a and c). For multiple-choice items the questionable assumption of random guessing is usually made.

Lord (1970 b) is an influential foundation paper the outcome of which is somewhat pessimistic for tailored testing. The pessimism is attributable to the low value of parameter a which he largely assumes (a value of 0·5 - corresponding to a point biserial less than 0·3, see Figure 7) and the constraint of a test of fixed length which he works within. (Green (1970) provides a healthy counterblast keeping variable test length in mind as illustrated in Figure 3.) The variables Lord investigated include,

(i) up-and-down step size, that is the fixed difference in difficulty (parameter b) between adjacent questions.

(ii) the value of a smaller up than down step where a probability of chance success exists. This is referred to as offset.

(iii) the method of scoring. Some possibilities would be

- the average difficulty of items attempted, excluding the first (as common to everyone) but including a notional (n+1)th item that depends on performance on the nth, final, item.
- the final difficulty level, that is of the (n+1)th item as in Figure 5.
- the conventional number-right score.

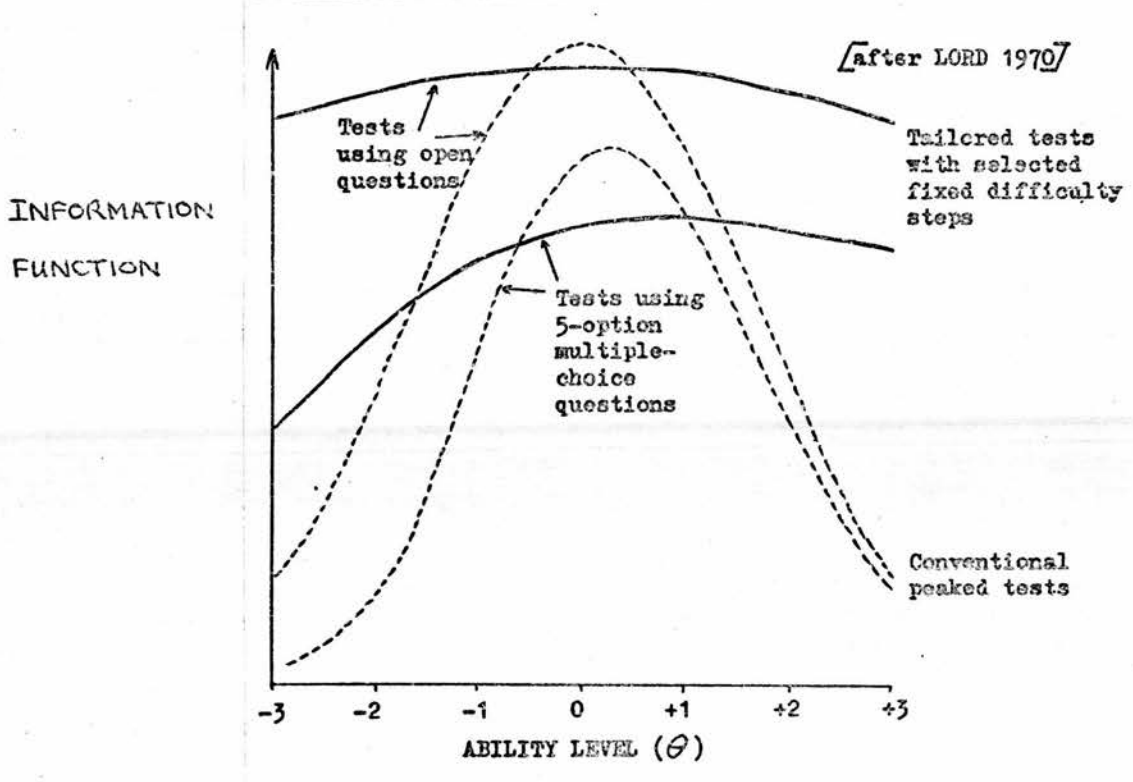
(iv) the effect of chance success.

(v) the value of Robbins-Monro shrinking step procedures (introduced earlier, in Section A).

Figure 11 illustrates some typical results. The tailored tests (solid curves) are more effective at the ability extremes, the peaked tests (dashed curves) more at the central ability at which they are aimed. The probability of chance success depresses the information function and leads to asymmetry in both curves. Peaked tests are idealised

FIGURE 11 Typical up-and-down branching test results for open-ended questions (c=0) and multiple-choice questions (c=0·2) compared with conventional peaked tests.

(Theoretical results after Lord, 1970 b)



fictions; the semi- or pseudo-peaked test in applied existence would have a curve of intermediate shape which might or might not top the tailored test curve for central ability.

A paraphrase of Lord's further conclusions is,

(i) the number-right score is perfectly correlated with the final difficulty score.

(ii) in terms of the information function the average difficulty score provides better measurement (and this score is subsequently used).

(iii) for 60 items a step size of 0·4 (in the difficulty level parameter b) seems best, and for 10 items a step size of 1·0.

(iv) when there is chance success offset step sizes improve accuracy of measurement.

Stocking (1969) essentially followed Lord's study but for a 15-item branching test. Her conclusions also followed Lord's but she was also able to study Robbins-Monro shrinking step procedures more extensively. These were found to be marginally superior to the best fixed step procedures. However, for an n-item test the Robbins-Monro procedure calls for 2^n-1 items - over 32,000 items for a 15-item test. Hybrid procedures were studied which attempted to capture some shrinking step advantages using a change in fixed step size, but the procedures tried failed to do so. Lord (1971 a) reaches the same conclusions for Robbins-Monro and hybrid procedures.

Mussio (1972) aimed to cut down the item requirements for a branching test by curtailing the item network at lower and upper difficulty limits. For example, a 60-item test restricted to 11 difficulty levels requires 605 items compared with 1830 for a full network. The penalty is some loss of precision at extreme abilities, but results

remain superior to those for a conventional test.

Further theoretical work for the US Army by Waters & Bayroff (1971) had the particular merit of looking at the effect of varying item discrimination. They compared various 5-, 10-, and 15-item branching tests with various conventional tests of the same length. Scores were evaluated by their correlation with underlying ability (after Lord (1952)). For item discrimination at 0·6 or above (as assessed by biserial correlations with underlying ability, not with a fallible total score for which the equivalent values would be lower) the highest correlation was always for a branching test. For lower item discriminations a conventional test achieved equivalent results, while for the lowest biserial assumed, 0·3, a conventional test was superior. This latter result can perhaps be regarded as an indicator of a conventional test's robustness under conditions of misuse. All the observed differences were small, perhaps expectedly so for a global measure like the correlation coefficient. To round off the series of US Army studies it is appropriate to mention here the work reported by Bayroff, Ross & Fischl (1974). Here they describe a sophisticated online individualised test set-up - more sophisticated than the equipment visualised by Bayroff (1964). Cynically this appears a case of the electronic technology overtaking psychometric technique for they report no plans or decisions for the forms of individualised testing to be tried. Essentially this might be taken as a comment that it is not yet clear that any form of tailored testing has established a convincing case.

Finally on branching test research an all too rare empirical study is reported by Larkin & Weiss (1974). They worked with multiple-choice vocabulary items. Three 15-item branching tests were used and a variety of scoring methods. Both the branching tests and a 40-item

pseudo-peaked conventional test were administered by online VDU. Three groups of over 100 students each took one of two of the branching tests. Two groups also took the conventional test. All groups were retested after 5-10 weeks, two on the same branching test.

The average difficulty score consistently had the highest test-retest correlations (confirming a superiority shown in Lord (1970 b)). Test-retest coefficients for this method were of the order of 0·86. In comparison with 15-item conventional subtests the testing design was such as to permit the disentanglement of memory effects from the test-retest stabilities. This is important because whereas, in a conventional test all items are repeated on retest, in an individualised test this is not so. In fact in the 15-item branching tests about 8 items were repeated on average. Taking the memory effect into account the branching tests showed the greater stability.

Intercorrelations among the various scoring methods were all high - always over 0·9 and often over 0·95. The correlation between average difficulty score and final difficulty score was 0·91. This is of interest because the latter scoring method is the one generally used in pre-1970 studies, while the former now seems clearly preferable. Some of the results of the earlier studies may have been a little more favourable had average difficulty scoring been used.

The theoretical branching test studies consistently demonstrate superiority over a peaked test outside the central ability range, possibly for a pseudo-peaked test this could be so across the whole range. The relationship with underlying ability is also a little closer. In an empirical study test-retest stability was a little higher. The tendency for the tailored approaches to nudge ahead is showing

more consistently here. Notwithstanding this the writer anticipates that research on branching tests will tend to decline in favour of the newer methods to be described at 6 and 7 below.

5. Flexilevel tests

The flexilevel test is an ingenious attempt by Lord (1971 b) to produce a practicable pencil-and-paper procedure with the capacity for a limited degree of tailoring. As such it is peripheral to this thesis. However, two empirical VDU administrations are reported in the literature and these will be described - in the view of the writer these attempts are misguided.

Consider a conventional test of 61 items arranged in item difficulty order from easiest to hardest. Item 31 will be at the centre of the difficulty order with 30 items easier and 30 harder. Imagine the 61-item test bent in two so that on the printed page item 31 is now uppermost at the head of two columns of items. One column on the left, say, is the easier set and will now be found in descending order of difficulty - items 30, 29, 28, and so on. The harder set in the right-hand column is items 32, 33, 34, and so on. In a flexilevel test the testee begins with the single item at the head of the page, that is with what was item 31. (For the flexilevel test the items will be renumbered.) If he answers correctly he goes on to the next unattempted item in the right-hand, harder, column, or if incorrectly to the next available easier item going down the left-hand column. Testing proceeds following this rule until, in this case, 31 items have been answered. Switching from column to column is inefficient (although it completely overcomes the danger of misrouting), but necessarily the 31 items attempted will tend to include that subset

most appropriate to the testee, and these will have been attempted in the course of 31 rather than 61 items. In this way limited tailoring is achieved. The method requires a self-scoring form of answer sheet which will indicate if an answer is right.

As a pencil-and-paper test the format is somewhat demanding. In a study with Eighth Grade pupils 10% of answer sheets had errors in applying the procedural rules. However, online VDU presentation overcomes the administrative problem. It might well be argued that the flexibility of online presentation is largely wasted on the inefficient flexilevel scheme (drawn up explicitly for the constraints of pencil-and-paper).

Hansen et al (1974) propose to use computer-based flexilevel testing in US Air Force technical training. The method has the advantage that it can use existing tests directly - although it would not be expected in this case that the limited tailoring would recover the full loss of reliability from reduced length. Betz & Weiss (1975) carried out both empirical and Monte Carlo simulation studies of flexilevel testing. The simulation was based on the characteristics of the same pool of multiple-choice vocabulary items used in the empirical study. A flexilevel test of 40 items was given together with a conventional pseudo-peaked test of the same length. In the empirical study both tests were administered by online VDU, 367 students taking the flexilevel test of whom 227 also took the conventional test. Some students were also retested. Test-retest stability coefficients were comparable for the two test forms at about 0·89. The parallel forms reliability from the simulation study was higher for the flexilevel test - a mean of 0·84 as against 0·80. Correlation with under-

lying ability was marginally higher for the flexilevel test, 0·91 as against 0·89. The simulation study was also able to look at the information function of the two test forms in relation to ability. Figure 12 summarizes some of the results. Being based on real item pools these curves are in substantial contrast to the peaked and flat crossing curves typical of conventional and tailored tests in theoretical studies (compare Figure 11). Some features of Figure 12 can be explained by some differences in discrimination in the items used for the two tests. Establishing exact comparability in empirical studies is very difficult.

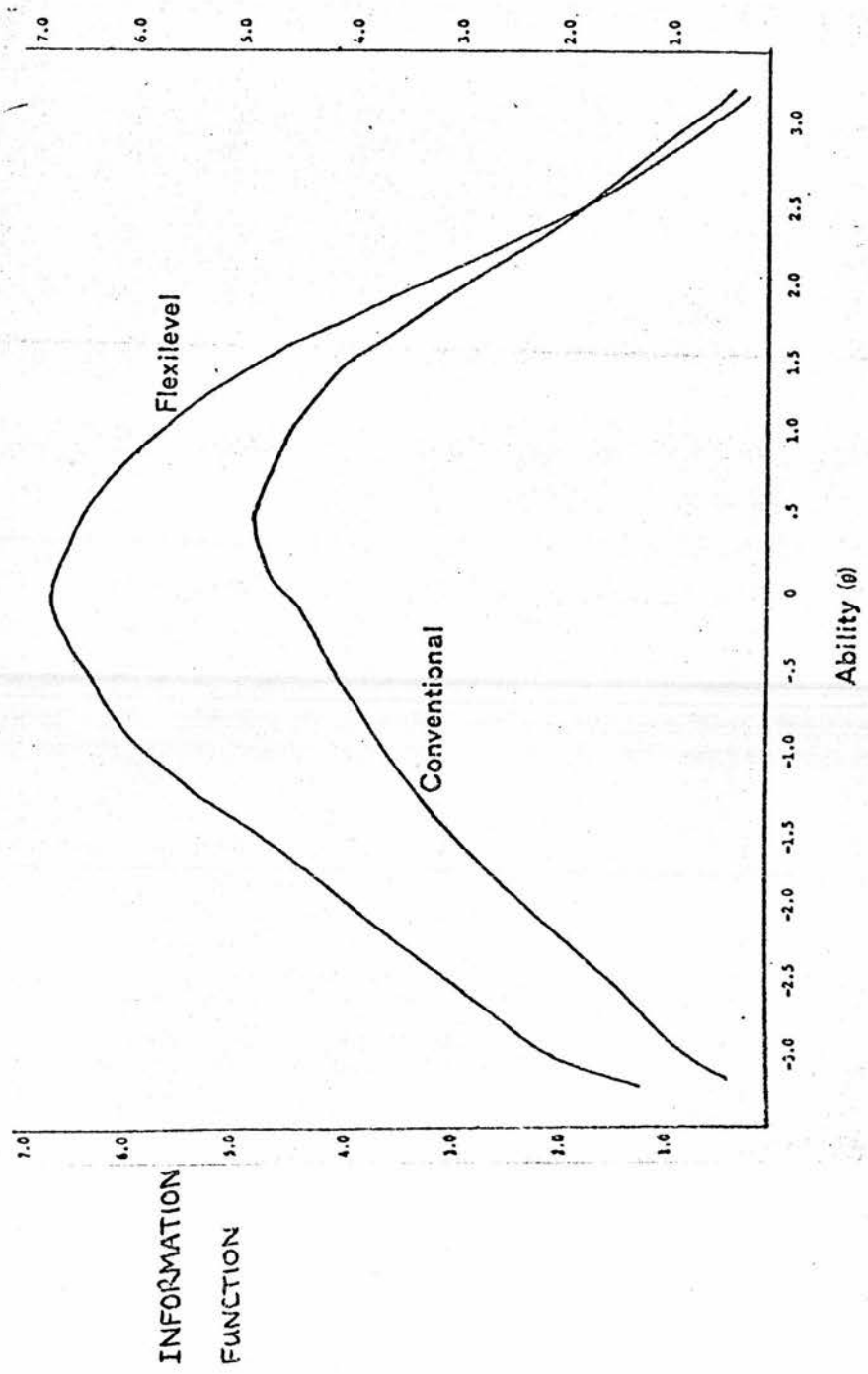
The simulation study here included a probability of chance success which was set at 0·2 because the multiple-choice questions had five options. The artificiality of this assumption is perhaps suggested by the correlation between flexilevel and conventional tests under empirical as compared with simulation conditions; this is 0·89 for the former (N=103) and 0·82 for the latter (N=10,000): however, there are also other factors making for consistency which affect real testees but not their simulations.

The stability coefficients from the empirical study (0·89) are the same as for the 40-item two-stage test of Betz & Weiss (1973) and only a little higher than for the much shorter 15-item branching test (0·86) using average difficulty scoring (Larkin & Weiss (1974)). The branching test can provide closer tailoring than the other two methods and the stability coefficients tend to confirm its greater efficiency.

In the approaches to test individualisation looked at so far a fairly consistent pattern of advantage (often small) over conventional testing has become apparent. The details of this advantage are

FIGURE 12 A comparison of simulated flexilevel and conventional tests based on real item pools.

(from Betz & Weiss, 1975)



confused by the difficulties of comparative empirical studies and by the simplifying assumptions necessary in theoretical work. Benefits in measurement precision at extreme abilities and a closer relationship to underlying ability would be conservative claims. However, all the methods are limited in the degree of tailoring they provide. In the cases of two-stage and flexilevel tests the limitation is their physical structure which limits the item pool and the number of possible routes. The limitation for branching tests is partly inherent in its item requirements, and the fact that in practice they will be imprecisely met, rather than in its physical form.

A concept of especial relevance to tailored testing (but not to conventional testing) is what might be called resistance to anomaly, or maintenance of equilibrium. The tailoring process can be viewed as a control or steering mechanism. The target is questions of appropriate difficulty - matching testee ability. The adjustments that have to be made to a testee's route must be sensitive to his current performance, but not over-sensitive, or else the occurence of anomalous responses will result in excessive reorientation and possible loss of bearing and consequent need for recovery. In engineering a servo-mechanism to correct the mismatch between course and direction is subject to damping so that wild movement or oscillation are avoided. The tailoring process needs to have similar damping to give it the required control characteristics. When a test is nicely on target an even balance of right and wrong answers will be produced with small variations in question difficulty.

Essentially in two-stage testing there is no damping nor recovery mechanism. There is one steering opportunity only. The routing test

aims the testee by dead reckoning and the course, once laid, is beyond further control. Lord has demonstrated that an appropriate multilevel test can absorb a certain amount of target error.

A flexilevel test has only two directions but it has a steering choice between fixed alternatives after each question: it has damping for the current direction but not for the alternative. A flexilevel test passes through the target questions and continues, it has no recovery after overshooting.

The small change in difficulty between the successive items of a branching test gives damping in both directions (easier and harder). Again there is a steering decision between fixed alternatives. The test can consequently hover in the target zone.

The amount of damping is important. The more damping there is the more items are needed and the slower the test is to reach the target zone. The Robbins- Monro shrinking-step procedures have an increasing damping as the test proceeds, but this also impairs their recovery after anomalous responses.

The approaches discussed so far have very limited steering or tailoring capacity. The procedures to be looked at in part 6 below differ in having much more flexible control over steering.

6. Item-finding procedures

Tailoring a test to suit a testee would be done most closely if each item were individually chosen rather than one of many predetermined item networks being followed. Ideally a procedure is wanted which at any stage in testing, after taking stock of the information to hand, will select the next item best to achieve the purpose of the assessment. Two such item-finding procedures have been proposed. The two

procedures differ in some points of approach and in the method of selecting the next item; they have in common a theoretical base in item characteristic curve theory.

Most work has been done on a Bayesian approach. Owen (1969) put forward a theoretical model which included the possibility of chance success. He assumes, and similar assumptions are common to most of the research discussed in this part, normal ogive item characteristic curves with known item parameters, and a normal prior distribution of testee ability. He derives an expression for the posterior distribution of testee ability that will obtain after answering a given question. He goes on to indicate a criterion by which that next question can be selected from the available pool so as to give the smallest variance in the resulting estimate of ability.

Owen's procedure includes two approximations. The posterior distribution always depends on a normal approximation for the actual distribution of ability prior to the current item. And in choosing the next item this is only optimum in the 1-step sense. How well a series of locally optimum single steps produces a globally optimum sequence is an open question.

Urry (1971 a) and Wood (1971) both used Owen's model. Urry carried out Monte Carlo simulations using three item banks, two of which were idealised, while the third took the parameters of an existing test. For his idealised item banks he took high values of item discrimination (a=1·6) with a probability of chance success of 0·2 - so that Figure 7 applies. 50 testees were simulated for each of these banks, and 100 testees for the existing-test simulation.

A distinctive advantage of item-finding procedures is that the

individualised test does not have a fixed length. As few or as many items may be selected in turn as are necessary to achieve a specified degree of precision. Urry specified standard errors of measurement of 0·32 and 0·25 as termination values (the assumed distribution of ability being taken as having a standard deviation of unity). In the case of the existing-test simulation (with items of lower discrimination) an alternative termination criterion of 30 items was additionally employed.

The less precise termination criterion was achieved by the idealised high discrimination item banks in about 11 or 12 items on the average, the more precise criterion in about 17 or 18. The existing-test simulation used an average of 27·5 items before reaching either the 0·25 precision criterion or the 30-item limit. Correlations with underlying ability were of the order of 0·94/0·95 - which is to be expected being only an alternative way of defining precision, although Urry presents this confusingly as a validity rather than a reliability relationship.

Even for the existing-test simulation these are good results. The reliability achieved in 27·5 items was comparable with that for the simulated test total score based on 80 items.

Wood's research included a Monte Carlo simulation based on a real pool of vocabulary items. Applying Owen's model he found that about 40 items were able to match a 60-item conventional test. Better reduction in measurement error was achieved in some parts of the ability range than in others and this could be attributed to the skewed nature of the item pool. A 60-item two-stage procedure was better than the Bayesian item-finding approach at the poorer end of

the item pool. In the Bayesian approach rapidly diminishing returns were experienced after about item 20.

Jensema (1974) also developed and tried out a Bayesian approach, again with minimisation of posterior variance as the criterion for successive item selections. A real data simulation was carried out using a response bank obtained from the administration of four quantitative tests to high school students. A sample of 5,000 pupils was used to estimate approximately the characteristic curve parameters of the items. From the 110 initial items sixteen were dropped as being too often unattempted, and a further 35 items were dropped as their discrimination (parameter a) was below 0·6. A further sample of 1,000 pupils was then used to obtain more exact maximum-likelihood estimates for the characteristic curve parameters of the remaining 59 items. At this stage one further item was deleted and 6% of pupils eliminated as repeatedly not converging during maximum-likelihood estimation.

The termination criteria were those of Urry's - a standard error of measurement of 0·25 or 30 items. The average number of items used was about 27. The ability estimates correlated 0·85 with the conventional 110-item combined test score, but this is inflated by a part/whole relationship and hence is surprisingly low.

Jensema also carried out Monte Carlo simulations using idealised item banks with item discrimination, parameter a, set at 0·8, 1·6, and 2·4. Estimates correlated 0·95 with underlying ability. For the least discriminating item bank no test sequence reached the required precision in 30 items - 35 items was a subjective estimate of the average number of items required. The two item banks with high

discrimination required an average of about 18 and 10 items respectively. Owen (1975) has produced a further theoretical Bayesian model which has the considerable advantage that it does not require an exact choice of item parameters. In practice item parameters will not be known exactly so that some tolerance is necessary. In other assumptions and approximations the model is the same as his earlier one.

Some of the Bayesian approaches have included a choice of starting point where there has been prior information to base this on. The capacity for a tailored start has generally been seen as desirable and likely to improve test effectiveness. However, Jensema (1974) also studied the value of prior information. Where prior information correlating 0·6 with the ability being assessed was available this only gave an average saving of about one item for the a=1·6 item bank. The saving would be greater for less discriminating items or for a less precise termination criterion. While any saving is worth having if readily available the value of an appropriate start is perhaps better viewed as largely motivational.

The limited evidence available on Bayesian item-finding procedures suggests an appreciable advantage over previous individualised testing approaches. Control over error of measurement is also a useful benefit. The procedures make a number of assumptions which will need further real-data simulation and also empirical studies to bring to light any resulting deficiences.

Despite the writer having argued that global item parameters can only be a first approximation to their usefulness in tailored testing, it is clear in the Bayesian studies that higher values of global item discrimination mean fewer items needed to termination. This is

because we are dealing here with assumed normal ogive characteristic curves. Given this theoretical basis three item parameters completely specify the item. However, real items can be expected to show some deviations from the assumed distribution. This will degrade the effectiveness of item selection, and the sensitivity of the procedures here to variations in item discrimination confirms the likelihood of this. Procedures which are aware of actual characteristic curves should show to advantage. Jensema (1974) makes a related point (p. 44), "A more basic question, which directly challenges the assumptions of the Bayesian item-finding model, is whether the guessing parameter is constant over all levels of ability. The model assumes that the Cg [guessing parameter] value is the same for any θ [ability] value. This seems questionable because an incorrect choice which appears reasonable at one level of knowledge may appear absurd at another."

It is characteristic of the Bayesian methods that before each item selection they scan all of the unused item pool. This requires much greater computing capacity than methods previously looked at. It also acts to limit the size of the item pool used, and this is a considerable snag as efficiency of testing would be expected to be related to the quality, the coverage and the depth of cover of the item pool.

The other proposal for item-finding procedures uses maximum-likelihood methods. That is, after any sequence of item responses it is possible, given known item parameters and assuming some form for the characteristic curves, to determine the ability at which the observed sequence is most likely. The item next selected is then the one with difficulty level closest to the current ability estimate.

Urry's (1970) is the only general research of this kind although Reckase (1974 a & b) also uses a maximum-likelihood approach working with the Rasch 1-parameter model which considers only differences in item difficulty.

Maximum-likelihood estimation can only sensibly begin once a testee has made both right and wrong answers. Consequently initial item paths have to be available to route the testee until he satisfies this precondition. Urry chooses to proceed immediately to the appropriate extreme of difficulty after a first question of median difficulty, while Reckase progresses by halving or doubling difficulty as appropriate until a contrary answer has been obtained.

Urry's Monte Carlo simulation study also had a basis in the Rasch model but went on to include a two-parameter variation which took the probability of chance success into account, and - more importantly - he systematically varied item discrimination. It is relevant to note that his approach, because of his initial routing, necessarily included items at one extreme or other of the difficulty range, for his results indicated that an item bank with a rectangular distribution of difficulty was better than one with a peaked distribution. This can be seen to be a direct consequence of his approach. He found for his method that item discrimination needed to be high, with parameter a at 0·8 or higher, to show advantage over conventional testing. When these conditions were satisfied considerable reductions in test length were achieved for the same standard error of measurement as compared with a conventional test.

The maximum-likelihood approach also requires an assumption for the form of the item characteristic curve, but does not require an

assumption for the distribution of ability. In scanning the remaining item pool to select the next question the specification is simpler than for the Bayesian approach being only a match on difficulty. On the other hand to update the ability estimate after successive items becomes increasingly onerous as all previous answers and their item characteristic curves must be appraised afresh. Again, then, the method requires substantial computing resources. In this case the computing requirements would act to limit test length rather than to limit the item pool. Urry's initial strategy in particular seems at risk to an anomalous response to the first item, but this is not central to the maximum-likelihood procedure and can be readily overcome.

Altogether the item-finding approaches show potentially high benefits, although the best results are achieved by unrealistically high levels of item discrimination. The requirement for high levels of global item discrimination, however, seems partly self-defeating. An apologist for conventional testing could justifiably argue that the conventional methods have evolved to work with the items that are available and that the availability of super-items is by no means guaranteed.

With continuing progress in computer technology there is perhaps little point in emphasising the possible restrictions from computer requirements. Even so there will presumably continue to be a cost advantage to methods which can function with slower smaller machines.

7. Stradaptive and Broad-range approaches

These two methods are due respectively to Weiss (1973 b) and Lord (1975 b); they have a number of similarities and can be regarded as simpler item-finding strategies. A stradaptive (from stratified

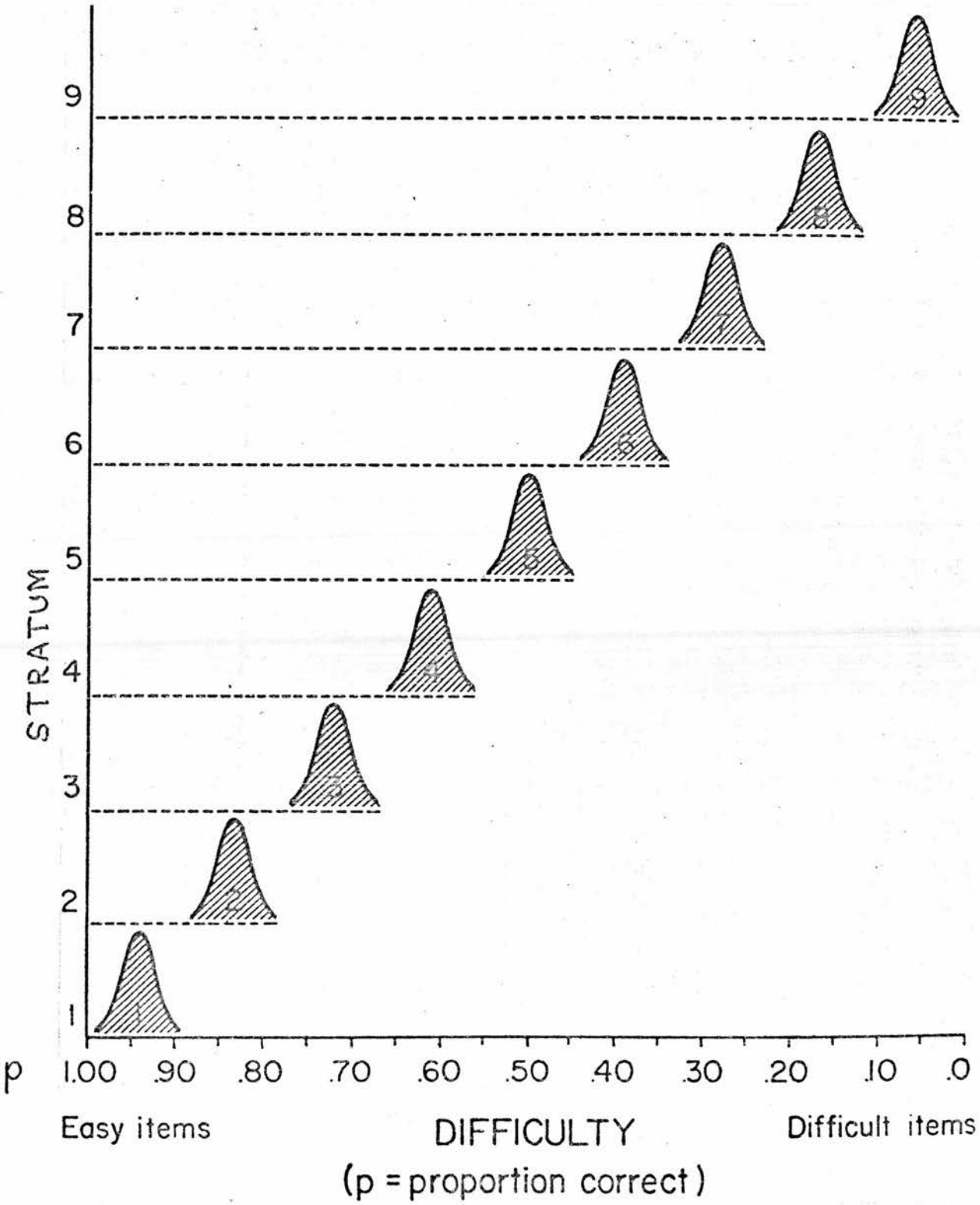
adaptive) test uses an item pool organised by difficulty level into a number of strata. Figure 13 illustrates the kind of distribution by difficulty that Weiss has in mind. All the items within a stratum are regarded as equivalent although they are queued for use with the most discriminating items first. A testee makes a tailored entry to the item pool at what is judged an appropriate stratum. Depending on his answer to the first question he is moved to a harder or easier stratum for his next question - the harder stratum following a right answer. Testing may continue for as long as required. Weiss draws an explicit analogy with a Binet-type individual test. He speaks of basal and ceiling strata, these being the difficulty levels at which success and failure are certain. Failure for multiple-choice items is taken as chance success - although the definition of this will be somewhat problematic and necessarily probabalistic. Several scoring methods and test termination criteria are possible. Weiss presents only illustrative results both here and in Weiss (1974).

A stradaptive test offers more controllable testing than all but the item-finding procedures. The reduction of an item pool to strata is a realistic device acknowledging the fallibility of the item descriptive information that will in practice be available. It has the deficiency as compared with the item-finding procedures that the method of test scoring does not directly yield an ability estimate - the associated advantage is that it makes no assumptions about item characteristic curves or the distribution of ability; at the moment its method of scoring is an open question. In forming the item strata only global estimates of item difficulty and discrimination are used.

Lord's broad-range tailored test is so called because its aim

FIGURE 13 An example of an item pool for a stradaptive test.

(after Weiss, 1973 b)

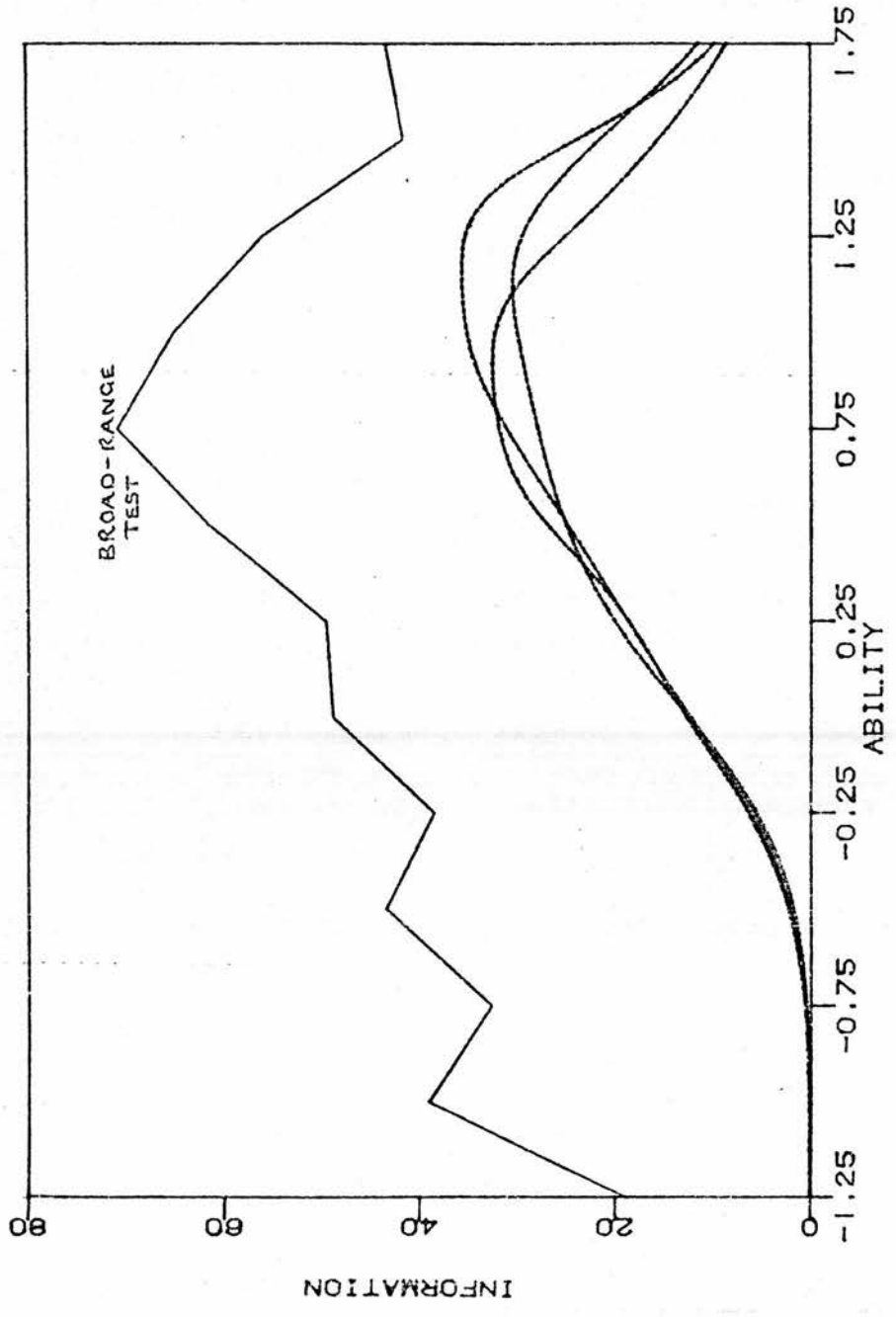


is to provide effective assessment from Fifth Grade pupils upwards. The test described is one of verbal ability: it draws on a wide range of existing tests to give an item pool of 182 items of five types. Items were chosen for type and difficulty level not for discriminating power.

The items are grouped in ten difficulty levels. Essentially items are again queued within difficulty level in decreasing order of their discrimination. In fact, because there are five item types certain adjustments of detail are made here and elsewhere to ensure a generally uniform mix of types. A testee makes a tailored start at an appropriate entry level. He is then routed to easier or harder levels depending on his answer. This routing continues only until at least one wrong and one right answer are available. At this point maximum-likelihood procedures are introduced in conjunction with item characteristic curve theory to find the ability at which the observed answers are most likely. Now the next item is selected, from all items of the appropriate type, that gives the most information at the estimated ability level. The procedure continues for a fixed test length of 25 items.

The design presented is reported by Lord as one chosen from about thirty following simulations based on 1,000 or so simulated testees. An item pool of double the size gave results that were twice as good, the gain being mainly attributed to the availability of more items and only partly to their arrangement in more and closer difficulty levels. Figure 14 shows the information function for the broad-range test (for entry at ability level 0·75) compared with that for three conventional tests adjusted down to the same 25-item length - the conventional tests are three forms of the Preliminary Scholastic Aptitude Test.

FIGURE 14 A comparison of a broad-range test and three conventional tests. (Simulation results from Lord (1975 b).)



Clearly the broad-level tests require greater computing capacity than the stradaptive test. The broad-range test has the advantages (and disadvantages) of a base in item characteristic curve theory. The stradaptive test on the other hand can be regarded as a flexible development of a branching test which could be administered - as Weiss points out - by relatively simple equipment. Lord does not give information about the kind of route through his 10-levels that is taken after his maximum-likelihood procedures come into play - granted that it becomes irrelevant to his method it would nevertheless be of interest if such routes approximated to some stepping rule.

The limitation of the broad-level test to 25 items is somewhat arbitrary, but some practical limit is imposed by the item pool available and possibly by the increasing computing load for maximum-likelihood estimation.

OVERVIEW

Sections A and B respectively reviewed the statistical antecedents of tailored testing and made a case for the context-free use of individual items.

It has been Sections C and D which have traced the development of tailored testing in educational and psychological measurement. Recent developments have shown considerable promise and there seems now little doubt that operationally useful instances of individualised testing will be with us shortly. However, there have been many difficulties - not least of which has been the translation from theoretical to empirical modes of research. A number of concepts and points have been picked up in the course of this chapter and these will provide

a framework for the method of tailored testing which is proposed in the next chapter.

3. A PROPOSAL FOR A TAILORED TESTING PROCEDURE.

A. Guiding Principles.

In devising a tailored testing procedure generally for selection and allocation and specifically for application in the Army a number of requirements were guiding principles.

1. Minimal item requirements

The procedure should be undemanding in its item pool requirements. While it is recognised that a better item pool makes possible a better test, it is important that testing should be possible with an item pool likely to be available. Once a testing programme is under way the extent and quality of its item pool can then be gradually increased with commensurate benefit: but if the minimum requirements are beyond what can be available - or if there exists the possibility that deterioration of an item pool in use could result in system failure - then that approach to tailored testing is unsuitable. What is wanted is a procedure that will operate with any item pool and produce a result at least on a par with what a conventional test drawing on the same pool could achieve. Such a procedure could be introduced with little practical difficulty and with confidence that it would be able to cope with fluctuations in the item pool.

The aim, then, is not for immediate operational excellence, but for some certainty of operational adequacy with potential for excellence -- this potential to be realised by long-term investment in the item pool.

2. Simplified item-finding

It almost follows from the emphasis on minimal item requirements

that procedures employing fixed item networks (such as branching tests) are ruled out. This is not so much because of the numbers of items involved as because of the tight item specifications to be met for assembling a network. It was, however, on the broader grounds of basic flexibility or agility that it was determined to go for a simplified item-finding procedure. An item-by-item choice is most in the spirit of individualisation, but more importantly this kind of flexibility was seen as necessary to give the procedure the robustness described. A fixed item network was seen as essentially more vulnerable under stress.

A simplified item-finding procedure (after the fashion of the stradaptive and broad range approaches of Chapter 2, D.7) was adopted for two reasons.

i. First was ease of access to the available items. Having to scan the full item pool before item selection is time consuming and would tend to require a larger more expensive computer. The necessity for scanning all individual items would act so as to limit the extent of the item pool and run counter to its progressive development. If on the other hand items are classified into a relatively small number of categories then access is by way of this fixed number of entry points and is unaffected by an increasing stock of items.

ii. the second reason is the imprecision of the descriptive item characteristics available. Selection from an item pool is on the basis of some statistic of item performance. To select on an individual item basis is to invest small differences in item statistics with greater precision than their estimation is due. Capitalising on small differences of little or no

validity if not leading to bias will neither be effective.

The item pool used here was classified into two sets of 19 bands on the basis of a novel index of item performance that is introduced later in this Chapter. The 19 bands are attainment[1] bands across the recruit population. The differentiation of 19 bands has the rationale that it is considered sufficient to meet the Army's allocation needs. To distinguish between classified items and an amorphous item pool the classified assembly will be referred to as an item library.

3. Explicit decision risks and direct interpretation

Cognitive assessments form only a part of the basis for Army selection and allocation decisions. By themselves such assessments provide an insufficient basis. However, reasonably valid cognitive assessments are more readily obtained than is the case for temperamental or motivational assessments, and each Army employment has its prescribed cognitive requirements. The cognitive assessments for a recruit eliminate from consideration many employments whose minimum requirements are not met. Allocation is then decided on motivation and the Army's needs. Later the causes of training wastage are substantially non-cognitive. In these circumstances it is more helpful to be aware of the risks of error attached to a decision on cognitive suitability than it is to have an estimate of exact cognitive standing. Wastage

---

1. This chapter presents the proposed tailored testing method using as examples instances from the data of the thesis. "Ability" is used in the general presentation as a generic term for what is being measured; the specific characteristic being measured in this research is verbal attainment and the examples given will refer to this. The data are fully described in Chapter 4: The Data Base.

during training is harmful to the individual trainee and costly to the Army. Awareness that the risk of failure through cognitive unsuitability is below some acceptable level leaves the selector with maximum freedom to attend to the influential non-cognitive factors.

In the first place decision risks in the testing procedure proposed will be in relation to placement on the existing test score scales. These are linked to performance in training as a result of the usual longitudinal predictive validity studies. Eventually it is a possibility that items could be directly calibrated in terms of criterion measures. In the meantime calibration is on the existing test scales.

4. Use of prior information and avoidance of assumptions

While by no means static Army selection and allocation has a great deal of stability. This status quo includes the applicant population and the range of employments. Any proposed testing procedure ought if possible to take advantage of this considerable body of knowledge. Information is available about the distribution of test scores for applicants in general, and also about the relationship between the first stage ACIO (Army Careers Information Office) assessments and those at the second stage selection centre. A recruit comes to a selection centre with an individual prediction available from his ACIO test scores. For tailored testing this means that a tailored starting point is possible.

A stable selection testing operation allows well based estimates of item performance to be obtained. The necessity for the assumption of particular forms of item characteristic curves can be avoided. For practical purposes items can be taken as performing as summarised in empirically determined characteristic curves. Similarly, as we

have seen, no assumptions need be made about the distribution of ability in the applicant group.

Working on a basis of empirically determined descriptions is not so neat and tidy as a sweeping theoretical approach but it has considerable advantage. Letting a situation display its local irregularities - which undoubtedly occur for reasons that are no less sound for their not being understood - avoids the mismatches that must follow a centrally imposed model. The use of local knowledge allows a messier but more efficient theory to be employed with ad hoc realities substituting for their more sweeping counterparts.

B The proposal

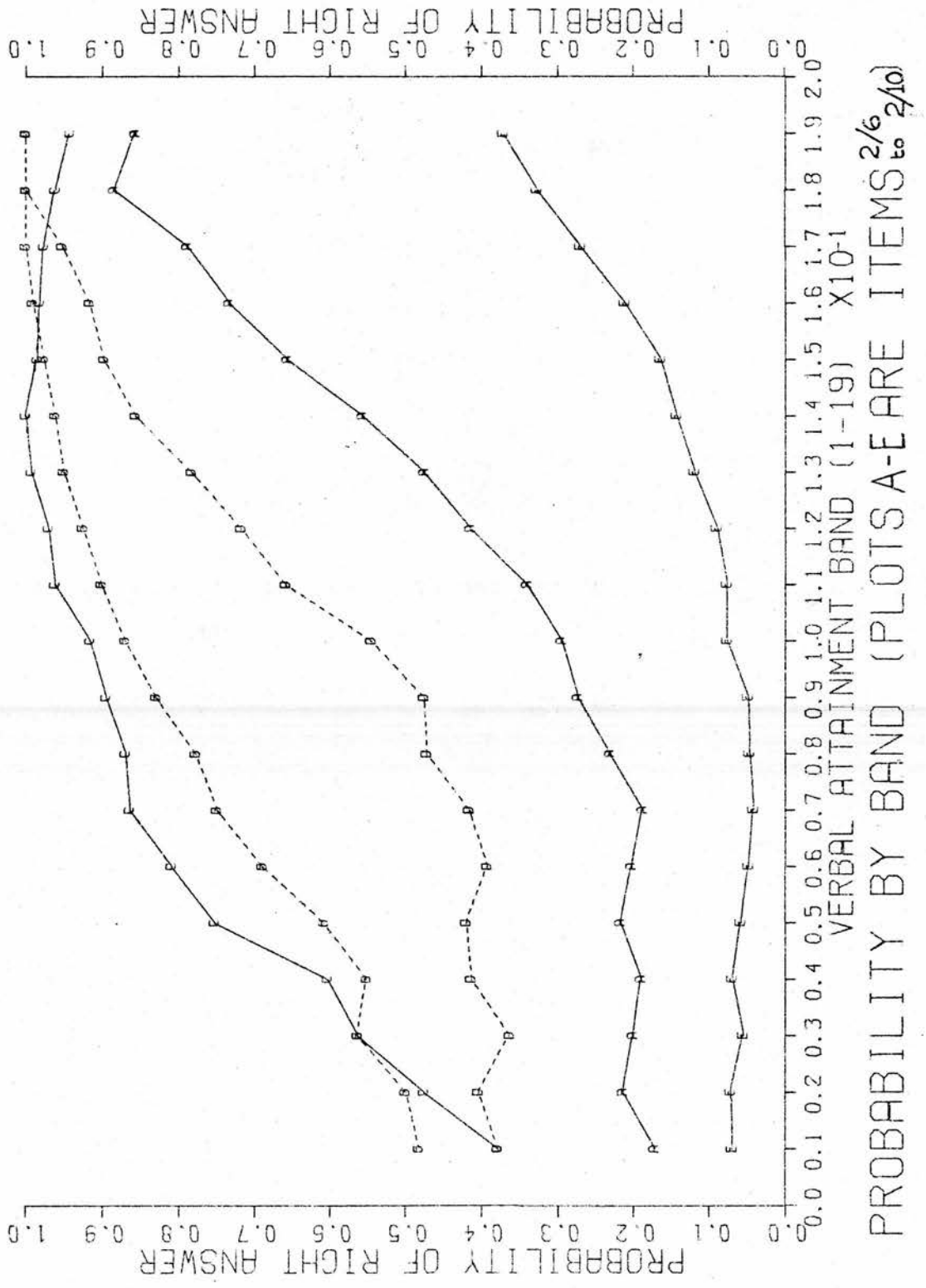
This Section describes in main detail the tailored testing procedure devised in accord with the guiding principles of the previous Section. Some more pernickety detail is in Chapter 5: Method, or is in Annexes referred to there.

1. Item-by-item ability estimates

Figure 15 illustrates an example of empirically determined item characteristic curves for five items from a verbal item pool from which the item library was subsequently selected. The item curves are labelled A to E. Curves B and D are dashed lines only to help distinguish crossing curves. The probability of answering each of these five questions correctly is plotted against level of verbal attainment. These curves are introduced fully in Chapter 4: The Data Base, but are mentioned here to help the description of the proposal.

Part of the prior information available is the distribution of verbal attainment in selection centre recruits. If for recruits with

FIGURE 15 Examples of empirically determined item characteristic curves for five verbal items from the pool used in the present study. (Note the scale factor of 1/10 applied to the horizontal axis.)



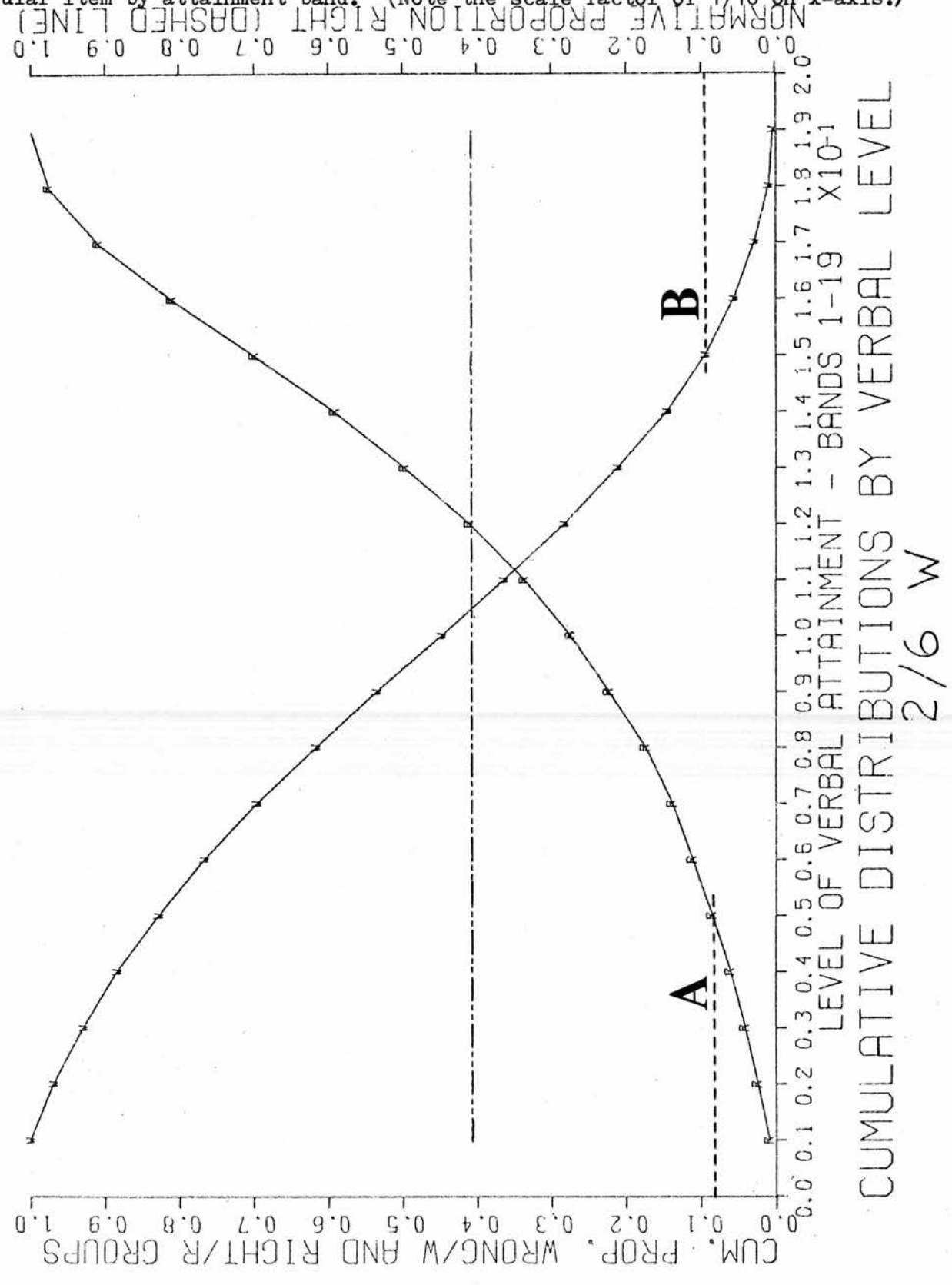
a given attainment the probability of their getting a particular question right (as illustrated in Figure 15) is multiplied by how many of them there are at that attainment band, and if this is done for the 19 bands, then the resulting products are the overall distribution by attainment of the recruits who get that item right. The complementary distribution for those getting the question wrong may be obtained similarly or by subtraction from the overall distribution. Wrong- and right-distributions are shown as the W- and R-curves in Figure 16 where the cumulative proportion of recruits is plotted against attainment band for a particular item[1]. This is item A of Figure 15 and is an item selected for the item library. It is a rather difficult item answered correctly by only about 40% of recruits - this is shown by the dashed line in relation to the right-hand axis.

The two curves of Figure 16 are cumulative in opposite directions, the R-curve indicating what proportion of the recruits giving right

---

1. Symbolic presentations are only given in the text here when they help clarity. Chapter 5: Method, will also present some such, but in particular it will also cross-reference to computer programs which necessarily give the definitive accounts (these programs, constructed by the writer, appear in the Annexes).
   For the substance of the paragraph referred to:-
   Let Da be the overall distribution of recruit attainment, and D is a discrete function, having values 1 to 19.
   If Pa is the conditional probability of item success on attainment level.
   Then,
   (DR)a the distribution of attainment in recruits giving right answers to the particular item
   is given by,
   $(DR)a = Da \times Pa$ for $a = 1,19$

FIGURE 16 Item/population derived distributions showing the cumulative proportions of recruits associated with success or failure on a particular item by attainment band. (Note the scale factor of 1/10 on x-axis.)



answers are accounted for up to and including any given attainment band, while the W-curve indicates the proportion from the total wrong group answering wrongly down to and including the given band. The reason for the opposite cumulations is for ease of inference as may be illustrated by the dotted lines at A and B in Figure 16. If a recruit's answer to the question is right then a decision that his verbal attainment is band 6 or higher would be in error for less than 10% of cases (line A). Similarly following a wrong answer the conclusion that attainment is band 14 or less would be in error for about 10% of cases (line B).

Cumulative curves such as those in Figure 16 are specific to a population as well as to the item. The distributions they depict can be referred to as item/population derived distributions. For simplicity the two curves will be called the wrong-curve and the right-curve.

A right answer is logically appropriate only for drawing inferences about the lower limit of the ability being estimated. The right-curve indicates what decision risk will accompany any particular placement of the lower limit on the ability scale. In practice too a right answer is also most useful for determining a lower limit and of little use at the upper limit: the upper end of the right-curve tends to become contiguous with the cumulative population curve (the more able people all tend to get the question right) so that right answers say little more about upper limits than could be taken from the upper limits of the population as a whole. The effect of a right answer is always uni-directional, it can act only to raise the ability level associated with any particular decision risk. If a person

gives only right answers then logically no inference can be made about the upper limit of his ability except through the constraint of the population distribution.

The converse of the last paragraph applies to wrong answers. The principle is entirely symmetrical, but quantitative asymmetry is introduced by multiple-choice questions for which there is the possibility of chance success. In Figure 15 for example, the probabilities of the easier questions top out at 1 while even the hardest question bottoms out at well above 0. This asymmetry acts to make right answers less informative than wrong answers. However, one aim of tailored testing is to match questions to the person. At the same time the view is maintained that guessing is essentially a response to questions which are excessively difficult for the testee. Consequently it is held, with some support from the research reviewed, that guessing will be relatively rare in tailored testing, and so right answers should not be substantially less informative.

Now, one right answer and one wrong answer will provide no more than widely separated lower and upper limits corresponding to some given decision risk, and what then? (For convenience limits that are associated with a specified decision risk will be called bounds.) The aim is to ask a sequence of questions so chosen as to be an approximate match to the testee's ability and, importantly, to yield an approximate balance of right and wrong answers. Assuming little guessing an equal balance is appropriate. The sequence of answers obtained has to be accumulated in some way after each answer so that lower and upper bounds converge until they are as close as required precision dictates - or, in relation to a specific selection decision, until the bounds

go above or below a particular selection cut-off.

To be able readily to accumulate the information from a series of answers requires the assumption of local independence. This is attributed to Lazarsfield (1959). It states that if a number of variables covary because of their joint dependence on a further variable then holding this latter constant will remove the covariation. The variables will then be independent within this local constancy of the further variable. In psychological measurement the assumption of local independence implies that for constant ability performances on items are uncorrelated. Answers made by one person fulfil the requirement for constant ability, and it then follows that the probability of success on two (or more) questions is simply the product of the separate probabilities for that ability level (these are the probabilities conditional on ability portrayed in Figure 15).

The assumption of local independence has been made in all the studies reviewed that have required the accumulation of separate item performances. As it was universal mention has been deferred till now where its rationale may perhaps seem more strongly relevant. It is a plausible assumption and fully in line with accepted concepts of relationships between variables, indeed assumption is perhaps too strong a term for an independence which follows from other basic postulates. It is interesting to consider what would follow if the assumption did not hold. Paraphrasing Lord (1971 e, p. 707), let us suppose that Pij, the probability of simultaneous success on items i and j, is not equal to Pi x Pj (as local independence would require) but is greater than this - where Pi and Pj are the conditional probabilities of success on items i and j taken singly (all probabilities for some fixed ability).

This would mean that some other psychological dimension is helping to determine whether items i and j are answered correctly. In other words items i and j constitute a two-item test measuring some psychological dimension other than ability. These items, and any others not exhibiting local independence, would be displaying heterogeneity of content. Hence if item libraries are homogeneous, and tests are aimed at single psychological dimensions, then local independence must follow.

No test of this assumption was found in the literature, perhaps because it amounts to a prescription for unidimensional tests and so merely endorses an accepted aim. However, it is necessary to check that it holds for any item library on which tailored testing is based, and such a check is made in Chapter 7: Results II.

Local independence can now be applied to the procedure being proposed. Following a person's answer to a first question the procedure has so far reached the item/population derived distribution for either a right or wrong answer. When the person now answers the next question local independence allows the conditional probabilities for this question to be applied in turn to the previous derived distribution (acting now as a prior distribution) to produce a new derived distribution. And so the procedure can continue, always taking the derived distribution following a previous question as the prior distribution for the next. Consider a sequence of questions i,j,k, and so on.

Let (DD)ai be the derived distribution of ability a following item i. Ability is taken as classified into discrete bands. (DD) is thus a discrete function and in the data of this thesis a has the integer values 1 to 19.

If (PD)aj is the prior distribution of ability before item j

Then,

$(PD)aj = (DD)ai$ for a = 1,19

Let Paj be the conditional probability for ability a of success on item j,

& let Qaj = 1-Paj

be a similar conditional probability for a wrong answer

Then,

$$\left.\begin{array}{l}(DD)aj = (PD)aj \times Paj, \text{ following a right answer} \\ \text{or} = (PD)aj \times Qaj, \text{ following a wrong answer}\end{array}\right\} \text{ for a = 1,19}$$

And continuing,

$(PD)ak = (DD)aj$

and so on.

After each item a new derived distribution is obtained from which new lower and upper bounds may be evaluated. One definition of how items should be selected for presentation in a tailored test sequence is that they should be effective in bringing about the convergence of the lower and upper bounds. The tailored test stops, as indicated earlier, when the required precision is achieved or when the testee is placed in relation to a specific selection cut-off.

2. Selection of library items

The criteria for selecting items for service in the item library are dependent on how the testing procedure uses the items. It is for this reason that this apparently initial task is dealt with after part 1.

Most individualised testing procedures choose the next item to try to match its difficulty to the current estimate of testee ability. The Bayesian item-finding procedure on the other hand looks for the item which will minimise the variance of the posterior distribution.

This principle is similar to the aim of the proposed procedure of converging the lower and upper bounds of the derived distribution. Two differences are,

i. that in translating a posterior distribution to a prior distribution for the next question Bayesian procedures have so far introduced a normal curve approximation for the prior distribution

ii. where the posterior distribution does not have a regular form the variance of the distribution is not readily usable for estimating decision risks at particular scale values.

However, other than through global item discrimination, no suggestion is made in the literature for choosing items inter alia for their convergence capability. This useful property is taken into account differently here. It was pointed out above that right answers act effectively to raise the lower bound of the current ability estimate, while wrong answers act to lower the upper bound. Information about the ability of individual items to move the lower and upper bounds is available in the data of the item/population derived distributions such as is illustrated for one item in Figure 16. It is the tails of the right-curve and wrong-curve (at which the lines A and B are drawn) that both set the bounds and also represent the interaction of an item with a prior distribution - and hence the tails determine the item's ability to influence the bounds. The item/population curves of Figure 16 are considered preferable to the item curves of Figure 15. This preference follows from the desire to make maximum use of prior information.

Each item has a right-curve and a wrong-curve and we cannot know in advance which will apply, but generally the questions which are

useful for moving one bound will be less useful at the other bound. It follows that questions should be asked with a particular answer in prospect and with some chance of that answer being given. The pursuit of converging bounds can be described as a search for the hardest questions a testee will answer correctly and the easiest questions he will get wrong. Right answers can be visualised as sweeping the derived distribution's lower extreme upwards, with wrong answers sweeping downwards at the upper extreme. Based on this rationale a requirement for a harder item during a tailored test will be met by reference to right-curve characteristics and vice versa.

There are two aspects of the tails of the right- and wrong-curves which appear relevant to their use as item indices:-

i. Tail location: the position of the tail on the ability scale will be related overall to the global item difficulty, but in particular it will relate to local difficulty at ability levels for which that tail of the item will be selected for use during testing.

ii. Tail discrimination: a tail will sweep more effectively (to use the metaphor of the previous paragraph) if item/population proportions cut off sharply with ability - graphically this is if the tail is blunter. For example, an item with a gradually tapering wrong-tail will tend to give a derived distribution with a long upper tail and so more widely separated bounds. This tail bluntness is called tail discrimination because it appears to focus on the ability of the item to be decisive in its tail region on the ability scale.

To be useful the concepts of Tail Location and Tail Discrimination have to be translated into quantitative indices for individual items.

The detailed indices will be given in Chapter 5: Method, but the principles will be introduced here. These are straightforward. For this first research use, indices with a simple direct link to the concepts were used. There is no claim that these are optimum indices of tail characteristics. The tails of distributions are notoriously among their less stable features: so the indices have to find what stability they can. Distribution tails are doubly important in the approach being taken as therein also lie the evaluations of decision risk. Any evaluation of decision risk can be no better than the accuracy of the tail information it incorporates or assumes.

An extreme percentile was taken as the index of Tail Location, say the 90th percentile. Using percentiles avoids susceptibility to the distribution of outlying values. On the other hand the index of Tail Discrimination should be susceptible to the distribution of outlying values and for this the absolute difference between a chosen percentile and the mean of the tail beyond that percentile was used. This is illustrated in Figure 17 where Ability P is the required percentile and M is the mean ability for the shaded area of the distribution tail.

Library items were selected jointly for range of Tail Location and high Tail Discrimination. Equal numbers of items were selected for their right and wrong tail characteristics.

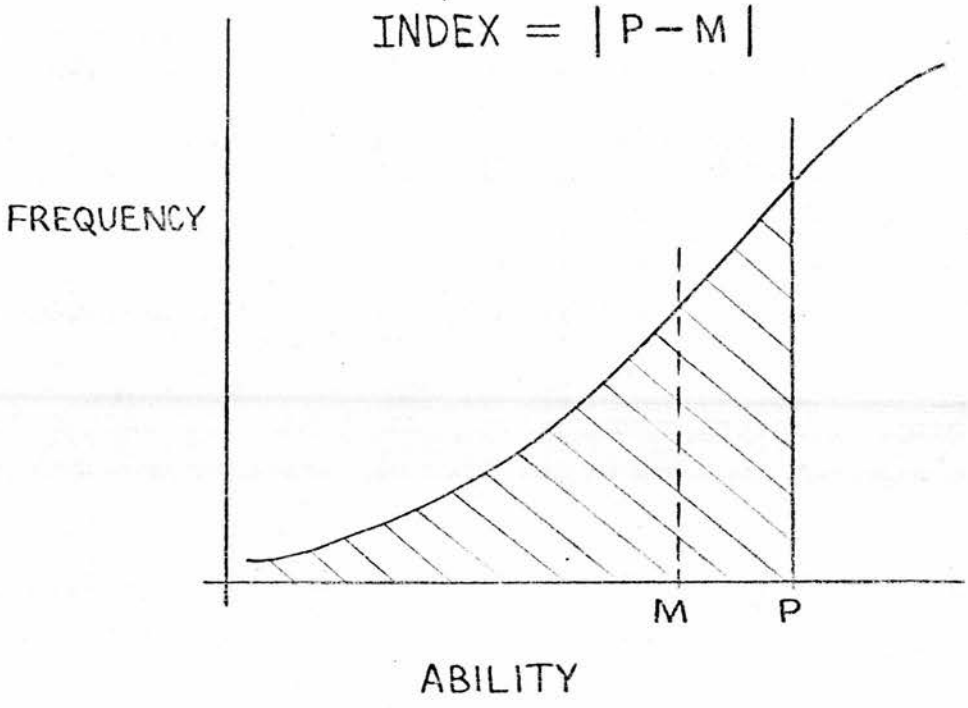
3. Conduct of a tailored test

In carrying out a tailored test there are three further kinds of decision to be taken for a testee,

i. the choice of the first question,

ii. the choice of the next question - repeated many times,

FIGURE 17 An index of Tail Discrimination.



iii. the decision when to stop testing.

To assist the choice of the first question advantage is taken of the relationship between the Army Careers Information Office (ACIO) test scores and subsequent scores at the Selection Centres which form the second stage of the Army entry procedure. A representative value is taken for the correlation between assessments at the two stages and this provides initial lower and upper bounds for a recruit's tailored test. The first question is selected on the basis of the tentative lower bound.

The relationship between assessments at the two stages could be capitalised on further by taking an ability distribution from the appropriate column of a well based joint scattergram as the prior distribution. This has not been done here because there are changes in ACIO testing in prospect which will result in a less satisfactory basis for such a prediction.

The best way of choosing the next item is an intriguing problem. The most usual solution has been to follow a wrong answer with an easier question and vice versa leaving only the amount by which the difficulty should change as a variable. Usually global estimates of item difficulty are used. The maximum-likelihood item-finding procedures match the next item to testee ability on parameter b, the difficulty parameter of item characteristic curve theory.

It is important for the proposed procedure to select items that will give an overall balance of right and wrong answers. This requirement stems from the logic of the situation and is not peculiar to this procedure. However, the wrong-then-easier or right-then-harder rule does not take advantage of the performance at

difficulty levels known from items earlier in the test. Consequently the rules for changing item difficulty described below include this element. And rather than use a global index of item difficulty the index of Tail Location was used which is at least localised on the ability scale to the general region where the appropriate bound will be. Item difficulty was deliberately expressed no more precisely than by attainment band.

As the tailored test proceeds two items of information will be noted,

i. the overall balance of right and wrong answers

ii. the balance of right and wrong answers at each difficulty level.

In relation to the difficulty of the question just answered the difficulty of the next question is given by the following simple algorithm using the information of i and ii.

	i. Overall balance		
ii. Balance at current level	More wrong	Equal balance	More right
More right or equal balance	0	0	+1
More wrong	-1	-1	0

Key
-1 move to an easier question
0 stay at same difficulty
+1 move to a harder question

As mentioned earlier an easier question is found using the Tail Location of the item/population wrong-curve and a harder question by way of the right-curve. The change in difficulty is by attainment band units and could be one or more.

The asymmetry in the table arises because in situations where alternative difficulties might be equally helpful the choice has consistently been to favour right answers. This is done partly because some predominance of right answers may offset any residual guessing, and partly - other things being equal - as a kindly, motivational gesture.

Finally the decision when to stop testing must be made. As indicated earlier this will normally be when the upper and lower bounds converge to the required precision. For the data to be presented here the required precision was defined as that of an independent 100-item conventional verbal attainment test for which results were also available.

C. Summary, Philosophy and Forward Glance

Essentially the tailored testing procedure that has been proposed is intended to be a coping procedure that will function in adversity. It is intended to function even with a relatively poor item pool. It adopts a simplified item-finding approach and a method of accumulating successive item performance which makes few assumptions and is explicit in relation to decision risks.

Items are selected for service in the item library partly for their tail-sweeping qualities in helping the derived distribution converge. This selection is carried out on the basis of an item's tail characteristics. Use is made of the known prior distribution of the characteristic being assessed.

In testing an individual person use is made of earlier individual information to make a tailored start to the test. Item selection

during testing aims to achieve an overall balance of right and wrong answers by manipulation of item difficulty using the position of the appropriate tail of the item/population derived distribution. Testing is planned to stop when measurement precision equals that of a 100-item conventional test.

The testing procedure makes no claim to be optimum of its kind. The procedure includes a number of novel approaches and the purpose of the remainder of the present thesis is to establish the fruitfulness of the approach. If the procedure fulfils its not unambitious aims then there will be a great deal of room for further sophistication. (However, these caveats are not an apologia for future results - which pleasantly surprised the writer by their general success.)

Having described the proposed procedure it will now be put into practice by a real-data simulation. Chapters 4 and 5 detail the data base used and the method of implementation. Chapter 8: Results III presents the simulated tailored testing results. Chapters 6 and 7 provide detailed information about the tail characteristics used and check the assumption of local independence on the data base.

4. THE DATA BASE.

Trials are called for to assess the value of the proposed tailored testing procedure and the concepts therein. Empirical trials are ruled out initially because of expense. A real-data simulation is indicated for the first evaluations. For its fidelity such a simulation depends heavily on the realism of the data it uses. The results of some of the research simulations of Chapter 2 suffered through a lack of realism. The meaning of realism here is that the data should arise under conditions that parallel as closely as possible the situation being simulated. However, the situations are generally different - otherwise why a simulation? - and some deficiences of realism can only be appreciated and not avoided. This Chapter describes the data base and presents its credentials.

A. Origins and description

Manpower Studies Section of the Army Personnel Research Establishment is engaged *inter alia* on test developmental research to replace the five tests of the standard battery used at selection centres. The field trials for this research are being carried out at the largest centre, the Recruit Selection Centre, Sutton Coldfield. This is the same Centre, which - because it is the busiest - could best support the introduction of individualised testing and which would benefit most from it. Data obtained at the Recruit Selection Centre thus comes from the same population of adult recruits as that for which tailored testing would first be implemented.

One of the five standard tests is of verbal attainment. This test uses items based on synonyms. There is good evidence for a

uni-dimensional verbal factor of this kind (a recent study is quoted below). Also this question type is among those for which the research reviewed in Chapter 2. B gave evidence of relative immunity from context effects. For these reasons data obtained in the course of trying out new verbal test questions for the replacement test is the chosen base for the real-data simulation[1].

The type of verbal item written for the new test was the common multiple-choice synonym or vocabulary item as exemplified here:-

Example	FAST	1. Drink
		2. Look
		3. Quick
		4. Time
		5. Vegetable

The recruit was instructed to find the word on the right that most nearly means the same as the single word on the left. In all 240 such items were written. During 1972/73 these items were pretested at the Recruit Selection Centre.

The method of pretesting was as follows. The 240 items were assembled as twelve 20-item tests. The twelve tests were administered in successive periods during the year, each for as many weeks as

---

1. The research project to produce a new standard selection battery is one which the writer was instrumental in setting up and for which he is Project Supervisor. Two junior colleagues, L.R. Preston and D.R.F. Hammond, have in succession been Project Leader and their ready help in making data available is gratefully acknowledged.

   The verbal test part of that project is reported in Killcross, Hammond, & Preston (in press).

necessary to accumulate a sample of between 300 and 400 recruits. From a recruit's viewpoint the tests were a part of the standard battery. As the purpose of the pretesting was to try out new items the tests were given without a time limit. Enough time was allowed so that all recruits could attempt all questions. In practice most items had 1% or less of omits, only 14 of the 240 items had more than 5% of omits with the maximum value less than 9%. Administration of the items under power conditions is also the best prescription for the use of the data in a tailored testing simulation. It is considered that at least the first empirical testing will take place under essentially power conditions, and it is under power conditions that item statistics have shown the greatest resistance to context effects.

Each item was tried out on an undifferentiated sample of recruits so there were many occasions when the less verbally able were confronted with difficult items. Such encounters make for guessing and chance success. It would be hoped that in tailored testing such mismatched confrontations would be avoided. Consequently guessing is likely to be more prevalent in the simulation data than in the event proper: this will have the effect of presenting more anomalous right answers to the tailoring process than it might normally have to cope with. It is hoped too that the simulation data will contain at least its fair share of anomalous errors: this is likely because the pretesting took place on the right population in the right setting and within the standard battery. A tailoring procedure has to be able to absorb anomalous responses, and if these are perhaps more prevalent in the simulation data then this provides a specially severe check on the procedure's ability. The philosophy adopted throughout in relation to the simulation is ungenerous; if in the simulation the procedure

has to cope with less favourable circumstances than in a real test then its results are unlikely to be misleadingly good - and the more likely it is that future empirical applications will find the procedure at least as effective as expectations based on the simulation results.

The detailed content of the 240 pretested items is immaterial to the present thesis but the nomenclature to refer to individual items has some relevance. This is based on the division of the items into twelve 20-item tests. The second item in the first test is item 1/2, and the last item in the twelfth test is item 12/20. (Figure 15 thus illustrates items 6 to 10 in test 2.)

During the 1972/73 pretesting at least 315 recruits took each of the twelve tests. The facsimile answer sheet at Figure 18 summarises the data collected from these test administrations. As well as the recruit's own recorded responses to the 20 items, Selection Centre staff later entered also his scores on the current verbal test of the standard battery (at the top right of the answer sheet).

Annex II gives the raw answer sheet data for the 4,472 recruit records obtained during administration of the twelve tests. The tabulations are described in full at the start of the Annex, and this practice will be followed for all Annexes of any complexity. Each record consists of a serial number, the test number, the recruit's score on the standard verbal test, the same verbal test score converted to a 1 to 19 attainment band (as will be given in Table 2 below), and his answers to the 20 items shown as option choices 1 to 5, or 0 for omit. The option choices can readily be converted to right/wrong data against the appropriate marking key. It is this right/wrong response data for particular items associated with independent

APRE ANSWER SHEET

NAME and INITIALS _____

NUMBER _____ DATE _____

TEST No

Cols 1 ---------- 6

For RSC use only

SP 103 : ARITH								SP125(A): VERBAL					
TOTAL		PART 1		PART2a		PART2b		TOTAL		PART 1		PART2	

Cols7 ---------- 20

EXAMPLES

A	B

1	2	3	4	5	6	7	8	9	10

Cols 21 ---------- 30

11	12	13	14	15	16	17	18	19	20

Cols 31 ---------- 40

FIGURE 18

Facsimile answer sheet.

estimates of verbal attainment (from the standard verbal test) that will constitute the ability/response bank for the simulation to draw upon.

The existence of the independent estimate of verbal attainment is a notable asset. The current verbal test from which this comes is a 100-item conventional test calling for guided responses to vocabulary items. The test is in two separately timed parts. In Part 1 a synonym is required that begins with three given letters, for example, GRIEF SOR....., (sorrow): in Part 2 the synonym must rhyme with a given word, for example, STOUT BAT, (fat, is the intended answer, meaning the same as STOUT and rhyming with BAT). The test has been in use for many years and normative data on recruit performance are available. A good estimate of the test's parallel forms reliability can be obtained from the correlation between Part 1 and 2 scores after uprating for the full test length. A recent sample of 963 recruits gave a Part 1/Part 2 product-moment correlation of 0·89, which gives a full test reliability (by the Spearman-Brown method) of 0·94. This is a high parallel forms value, especially as the item type varies between the two parts.

For reporting the outcome of an individual tailored test a coarser scale was required than the standard verbal test raw score scale. It was desired only to distinguish as many levels of verbal attainment as would satisfy the Army's need for differential allocation. Somewhat, but not entirely, arbitrarily 19 levels were taken. The accuracy of this value for allocation levels is of no consequence here. The conversion of raw score to attainment bands was as shown in Table 2.

Figure 19 shows the raw score and the attainment band distributions

Table 2 : The conversion of standard verbal test scores to attainment bands.

Raw score	Attainment Band
0 - 7	1
8 - 12	2
13 - 17	3
18 - 22	4
23 - 27	5
28 - 32	6
33 - 37	7
38 - 42	8
43 - 47	9
48 - 52	10
53 - 57	11
58 - 62	12
63 - 67	13
68 - 72	14
73 - 77	15
78 - 82	16
83 - 87	17
88 - 92	18
93 - 100	19

for 10,163 adult recruits passing through selection centres in 1973. This distribution is taken as the definitive population distribution for simulation purposes. Figure 20 shows the same distribution as a cumulative proportion and gives the mean and standard deviation in attainment band units.

B. Some supporting evidence on context effect

As pretesting was completed conventional item analyses of the 240 items were made using the external standard verbal test score as criterion. Subsequently a 100-item test assembled from the 240-item pool was tried out with a time limit at the Recruit Selection Centre. In this 100-item test the items were arranged in order of increasing difficulty and in general appeared in different item contexts than in the pre-testing phase - both in relation to specific neighbouring items and in relation to position within the total test. This provided an opportunity to look at the extent to which the item difficulties found in pretesting held up in the changed context of the 100-item (timed) test. This is relevant to the applicability of the ability/response bank to the changed context of tailored testing. Figure 21 plots the scattergram of the pretest and 100-item-test percentages correct for the 100 items. Item number in the 100-item test is also shown. For items in the first half of the test the relationship is very close. Subsequent items are affected by the speed element of the 100-item test and appear increasingly more difficult in this context. The evidence from the relatively unspeeded first half of the test makes a supporting case for the view that item performance under power conditions can be considered context-free in this instance also - this was

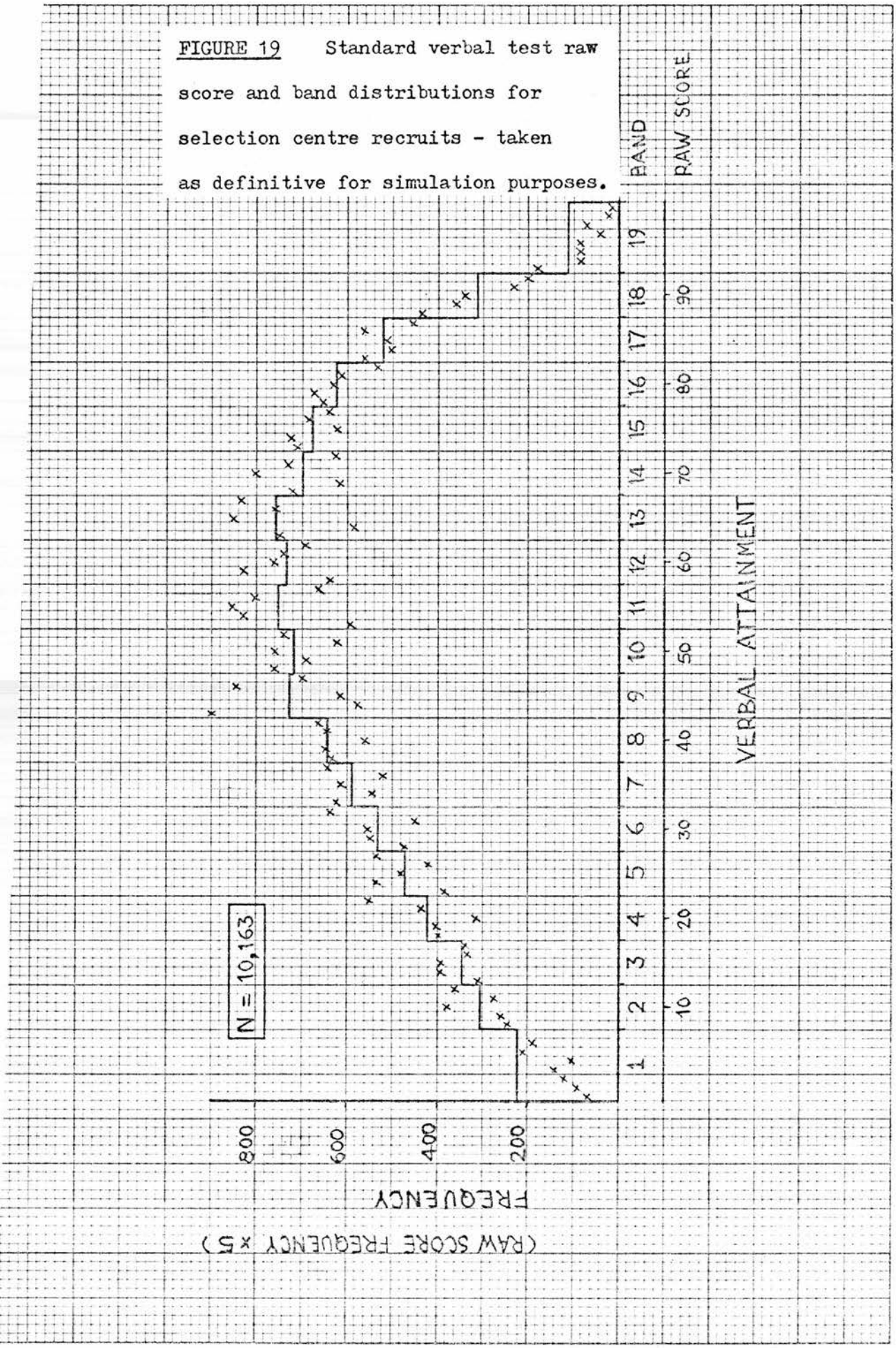


FIGURE 19 Standard verbal test raw score and band distributions for selection centre recruits - taken as definitive for simulation purposes.

FIGURE 20 Cumulative proportion distribution of standard verbal test bands for selection centre recruits.

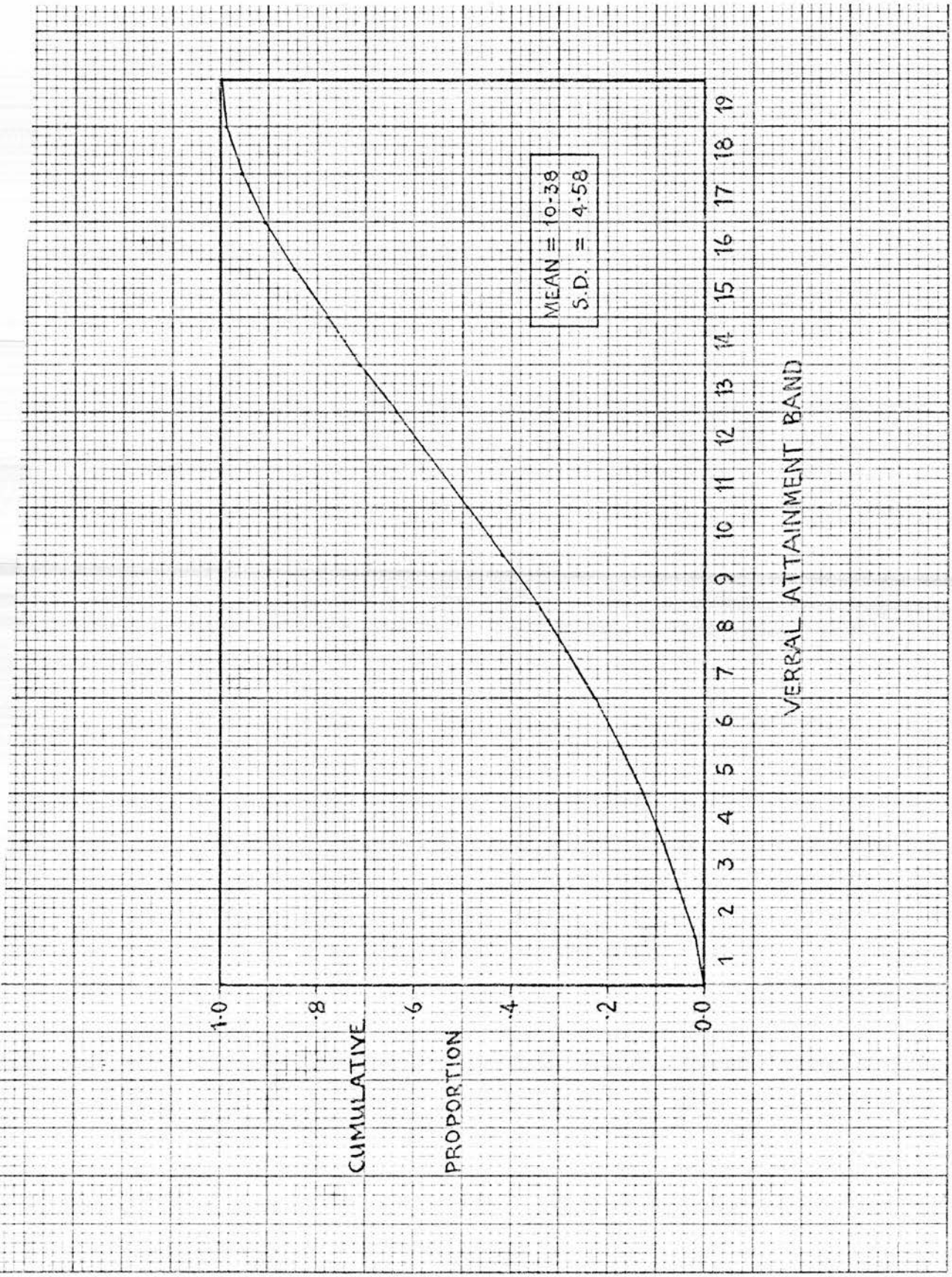
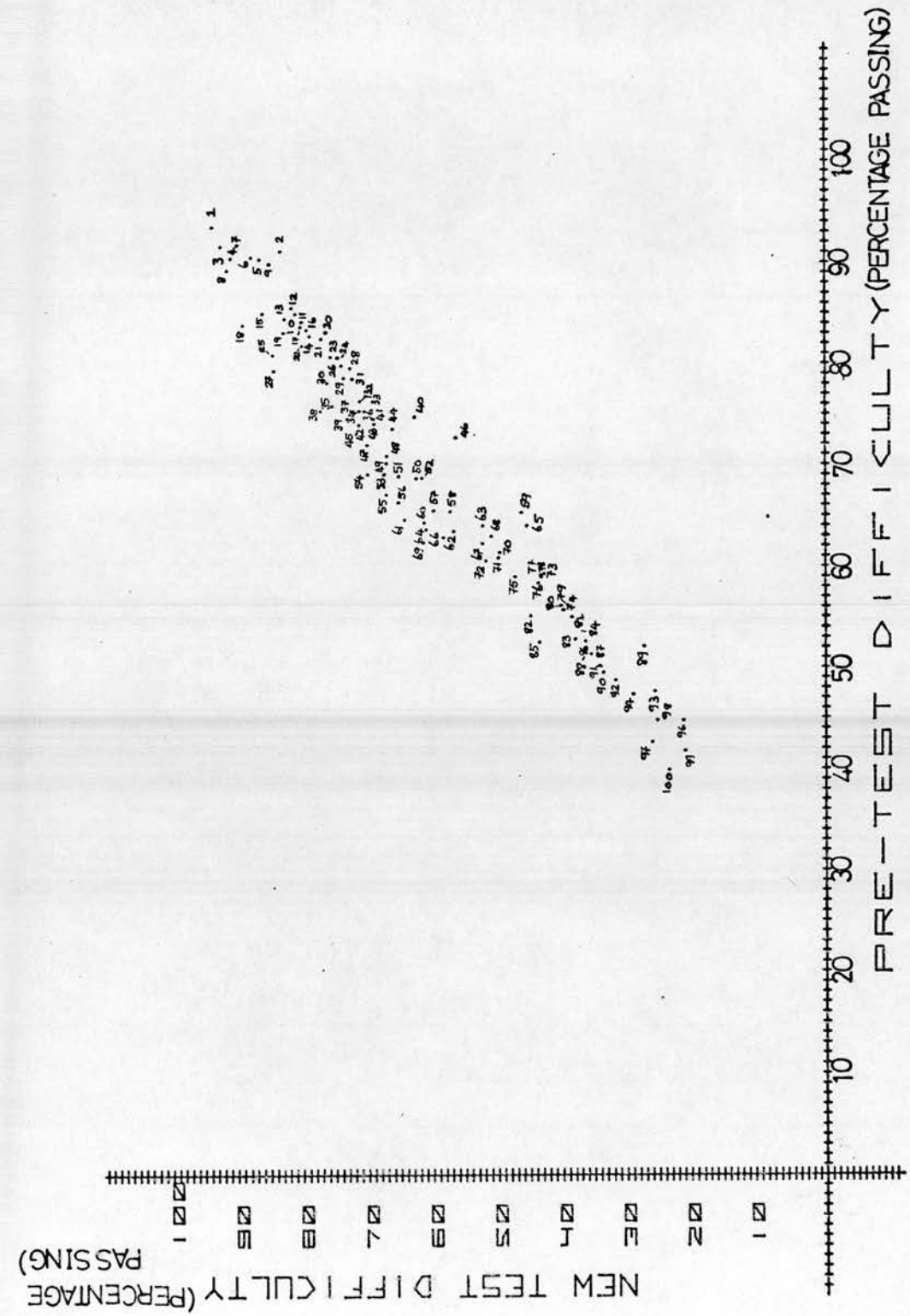


FIGURE 21 A comparison of the difficulty levels (indexed by percentage passing) of verbal items in two contexts - short, untimed pretests and a 100-item timed new test.



the consensus of a number of studies on pupil and student samples reviewed in Chapter 2. B.

C. A note on unidimensionality

The assumption of local independence has been shown to be equivalent to a homogeneity of content requirement. This assumption will be explicitly tested in Chapter 7: Results II, but some passing support can be appropriately given here.

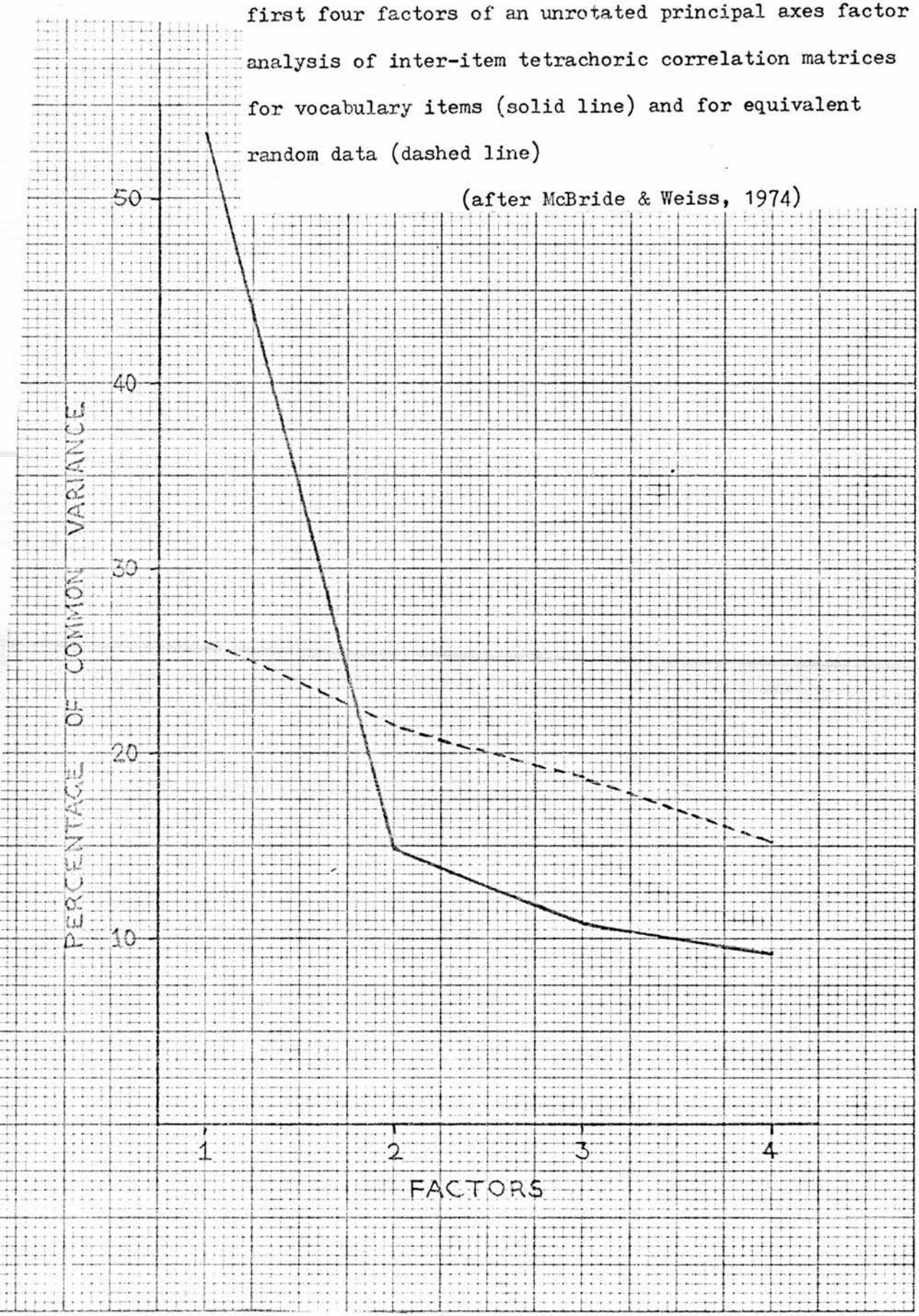
Because the 100-item test described in Section B is timed an estimate of its internal consistency reliability will give an inflated value. Consequently although the value obtained, 0·98, is high this must be offset by the degree of speeding that is visible in Figure 21. The amount of speeding is not excessive and some minimum support for homogeneity of content can probably be taken from the internal consistency reported.

More helpful is a study by McBride and Weiss (1974). They were developing an item pool for use in tailored testing, or in "adaptive ability measurement" to give their terminology. Their items were vocabulary items identical to the new multiple-choice items discussed above. They had a total pool of 575 items administered in various subsets - from 142 to 240 items - to undergraduate samples. The total student sample was, however, only about 500. 369 items were selected for future use and the unidimensionality of this selection was explicitly examined. Keeping up to a 10-to-1 ratio of testees to test items, six random samples of twenty items and their responses were drawn from the student records. An unrotated principal axes. factor analysis was carried out on each of the inter-item correlation matrices

from the six samples. Nine factors were specified. Equivalent random data matrices were similarly analysed to provide a base line. Figure 22 illustrates some of their results and shows the percentage of common variance accounted for by the first four factors. Only the first factor is above the random data base line. Their other results also strongly support unidimensionality. On a less selected group than undergraduates support for a single factor is likely to be even more striking.

FIGURE 22 The percentage of common variance accounted for by the first four factors of an unrotated principal axes factor analysis of inter-item tetrachoric correlation matrices for vocabulary items (solid line) and for equivalent random data (dashed line)

(after McBride & Weiss, 1974)



5. METHOD AND INTERMEDIATE RESULTS

This chapter gives details of how the ideas described in Chapter 3 have been investigated[1] using the data of Chapter 4. It is arranged in four sections:-

A. Deriving conditional probabilities from the raw data

B. Deriving tail characteristics and criteria for selecting the item library

C. Testing local independence in the item library

D. Real-data simulation of tailored testing using response banks from recruit/library-item encounters.

Sections A and B follow the method as far as the production of the intermediate results necessary for continuation to later sections. All the interpretation and final analyses will be made in Chapters 6 to 8 which will present and discuss the results from the methods of Sections B to D.

A. Deriving conditional probabilities from the raw data

The raw data of Annex II (introduced in the previous chapter) consist of answers given by recruits of known verbal attainment (as assessed by the existing standard verbal test) to questions from the 240-item pool. Each recruit record contains the twenty answers for one of twelve tests into which the pool was divided. These data are essentially from the

---

1. All the programming and data processing have been carried out using the computing facilities of the Royal Aircraft Establishment Farnborough, and in particular using online terminals running under GEORGE 3 and 4 on an ICL 1906S

answer sheet record shown at Figure 18.

First the test marking keys were used to convert the answer choices of the raw data to right/wrong form. Omitted answers were counted as wrong - this was appropriate as the tests were given under power conditions, and the proportion of omits was low in any case. The frequency of right and wrong answers was then analysed by attainment band. Attainment band was defined by verbal test raw score as given in Table 2 (p. 114).

Annex III gives the straightforward computer program, identified as PROGRAM 1, written to produce the required right/wrong frequencies for each attainment band.

Annex IV gives the required frequencies. The number of right answers is tabulated by attainment band for each of the 240 items. A detailed key to the tabulation is at the start of the Annex as usual. Within each of the twelve tests the items had been arranged in subjectively estimated order of difficulty and this pattern is discernible as well as that of increasing item success with increasing attainment band.

Now although the frequencies of Annex IV all arise from samples of over 300 recruits nevertheless after classification by attainment band many of the observed frequencies are small and would be expected to show considerable sampling variation. In particular, in an operational testing programme a much larger recruit base would soon be accumulated. The conditional probabilities (of item success by attainment band) from a large sample would tend to show gradual and progressive change of probability with attainment, whereas the frequencies of Annex IV would - as they stand - yield conditional probabilities a good deal

more irregular. The frequencies were therefore smoothed so as to improve the fidelity of the resulting conditional probabilities for use in the real-data simulation.

Before detailing the smoothing method used the distinction should be drawn between the smoothed frequencies and the true or large-sample frequencies for a particular item. The smoothed frequencies may be better estimates of the true frequencies than the Annex IV frequencies, although this is not necessarily so: however, the smoothed frequencies are realistic data for some plausible albeit hypothetical items similar to those tried and as such are better for the simulation than the unrealistic irregularity of the unsmoothed data. In so far as the smoothed frequencies will differ from the actual frequencies then, because the conditional probabilities will be linked with recruits' answers to particular items, this will result in some small deterioration in the efficiency of the simulated tailored test. This is in accord with the worst-case philosophy being applied to the simulation: this deterioration should not occur in operational testing.

The smoothing method was that of a moving average over a base of five attainment bands.

If $(F)b$ is the frequency of success at attainment band b for a particular item

& $(SF)b$ is the corresponding smoothed frequency,

then

$$(SF)b = ((F)b-2 + (F)b-1 + (F)b + (F)b+1 + (F)b+2)/5$$

where b takes the values 1 to 19.

At the extremities any frequencies for bands outside the range 1-19 are taken as zero.

If (SN)b is analogously the running average number of recruits in the bands from which the (SF)b arise,

then (SCP)b the smoothed conditional probability of item success for attainment band b,

is given by,

$$(SCP)b = (SF)b/(SN)b \qquad b = 1,19$$

Annex V gives PROGRAM 2 which takes in the frequency data of Annex IV and following the method outlined evaluates the smoothed conditional probabilities. These are presented in Annex VI. These are the data from which Figure 15 was plotted. By way of more expanded illustration Figures 23.1 to 23.8 display the conditional probabilities of the 40 items in tests 1 and 2 (items 1/1 to 1/20 and 2/1 to 2/20): the plots for the remaining 200 items of tests 3 to 12 are given in Annex VII.

In general the 240 items are well distributed in difficulty with some preponderance of easier items. The possibility of chance success on these multiple-choice items is well illustrated in the lower asymptotes of the curves for the harder items.

B. Deriving tail characteristics and the criteria for selecting the item library

The conditional probabilities of Section A (given at Annex VI) together with the known population distribution of verbal attainment (introduced in Chapter 4. A and displayed in Figure 19) allow the item/population derived distributions specified in Chapter 3 (at B.1 p. 95) to be obtained. In turn these derived distributions permit the evaluation of the item tail characteristics, Tail Location and Tail Discrimination, outlined in Chapter 3. B. 2.

FIGURE 23 (and continued in Annex VII)

Empirically determined item characteristic curves for the item pool.

FIGURES 23.1 to 23.8 follow.

FIGURE 23.1

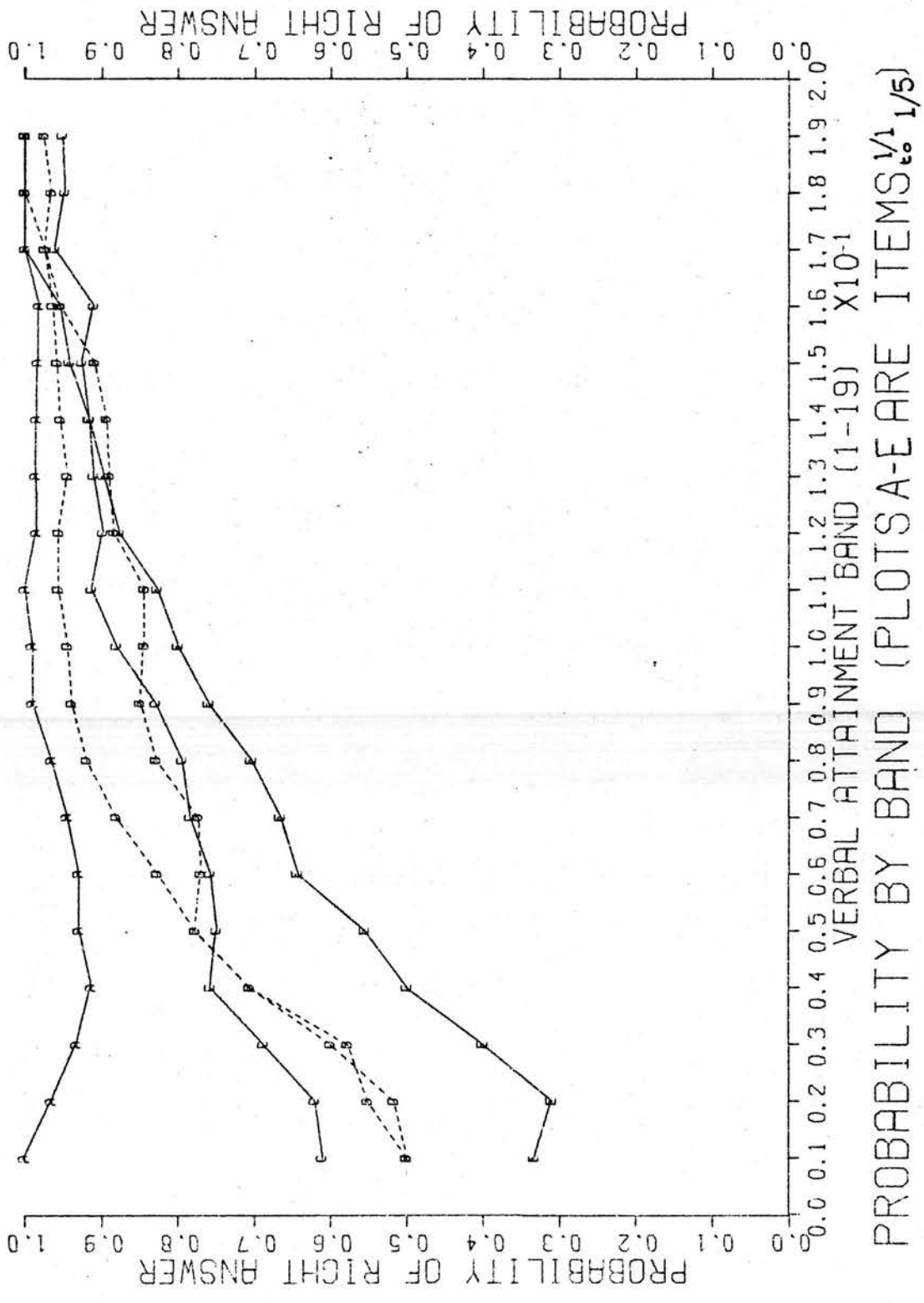

PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
VERBAL ATTAINMENT BAND (1-19) X10-1
PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 1/1 to 1/5)

FIGURE 23.2

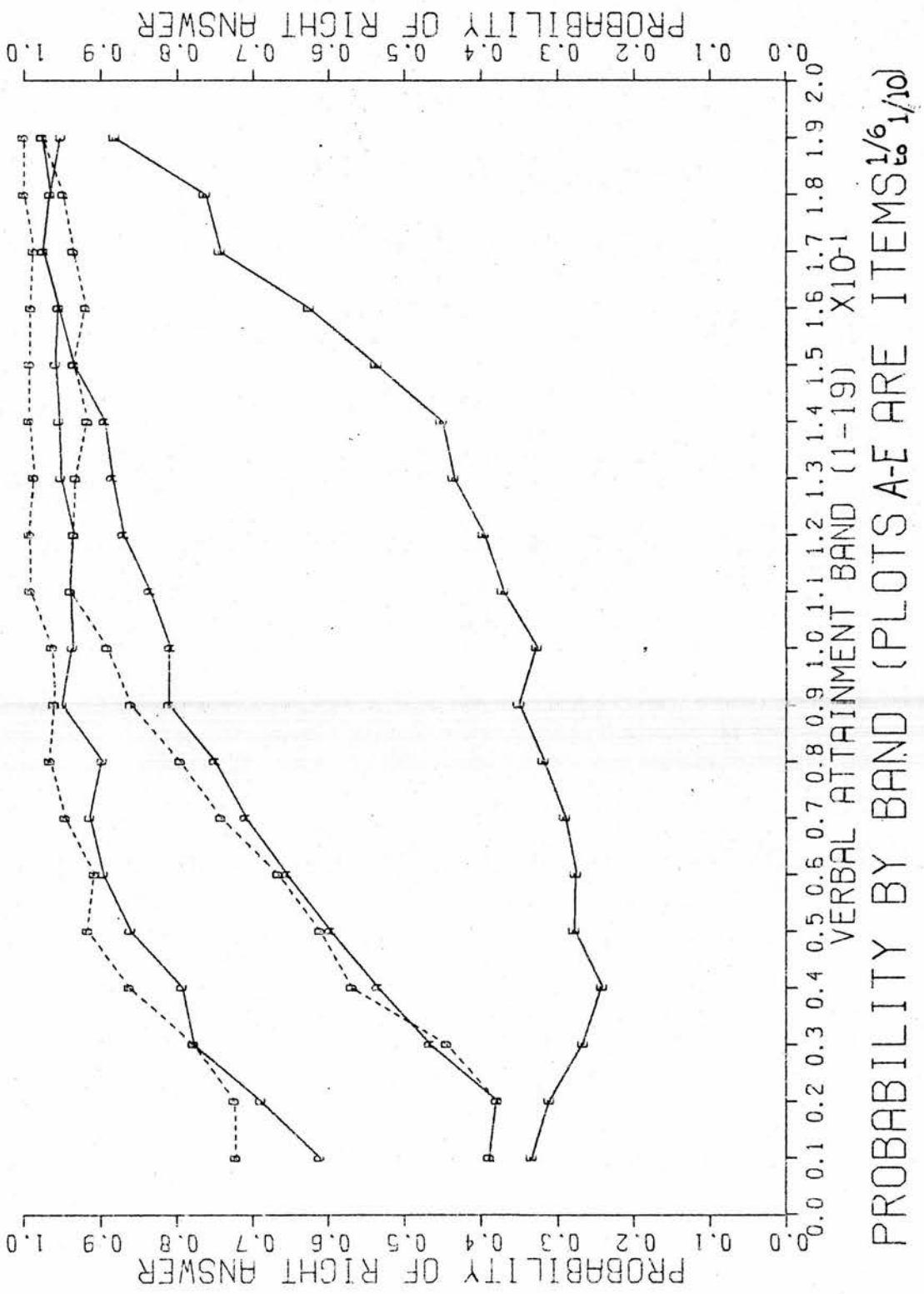

PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
VERBAL ATTAINMENT BAND (1-19) X10-1
PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 1/6 to 1/10)

FIGURE 23.3

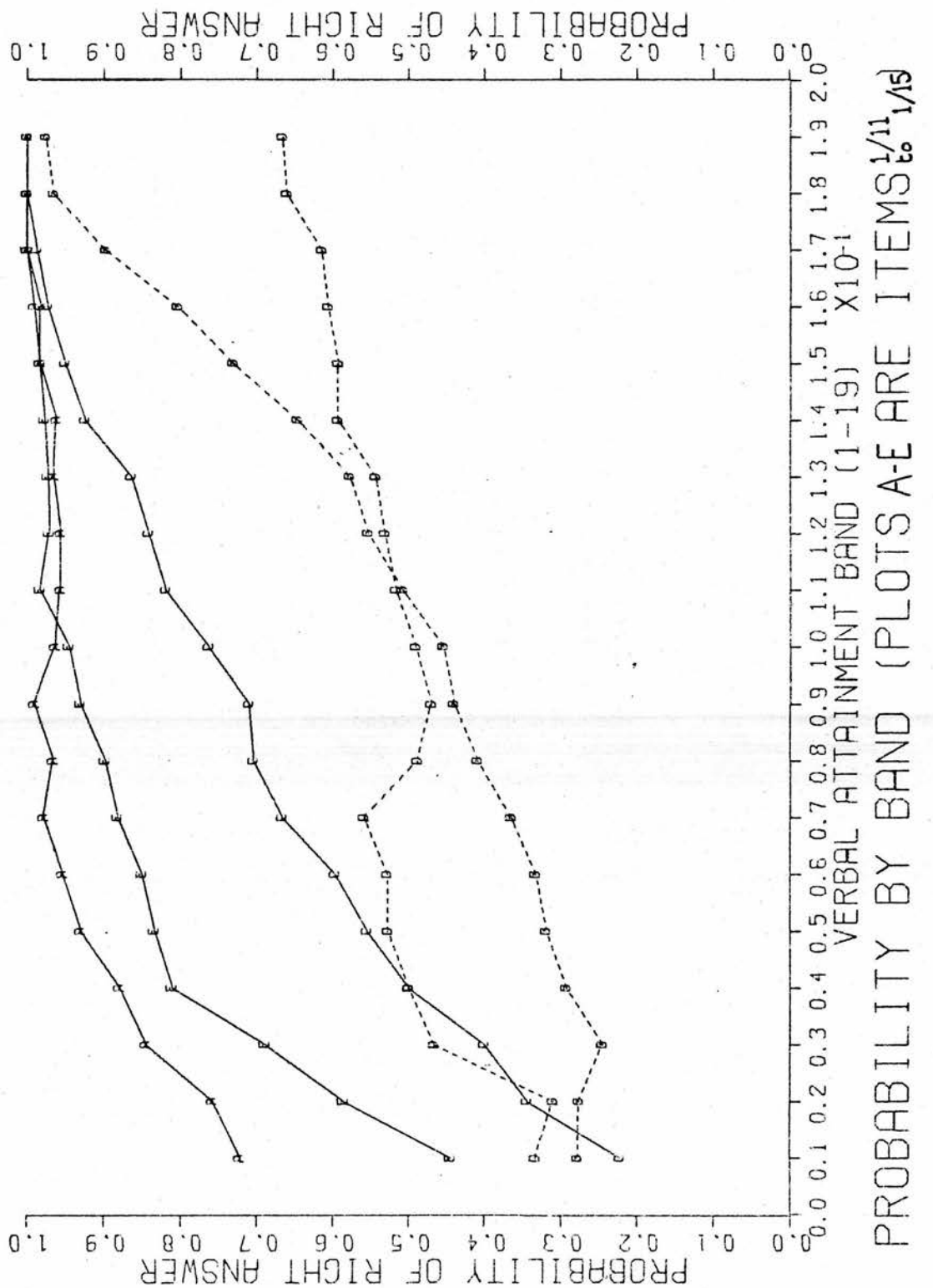

PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 1/11 to 1/15)
VERBAL ATTAINMENT BAND (1-19) X10-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 23.4

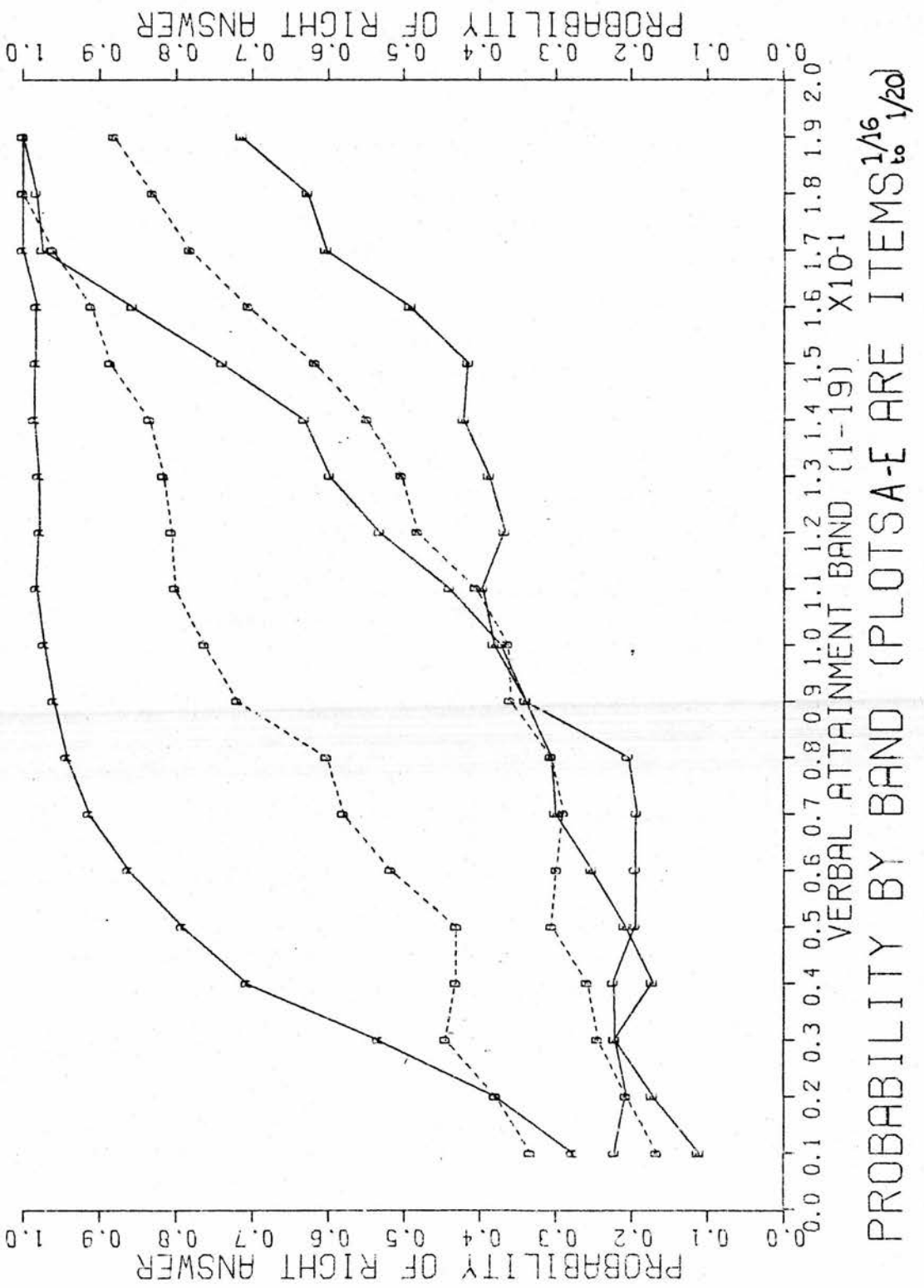

PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 1/16 to 1/20)
VERBAL ATTAINMENT BAND (1-19) X10-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 23.5

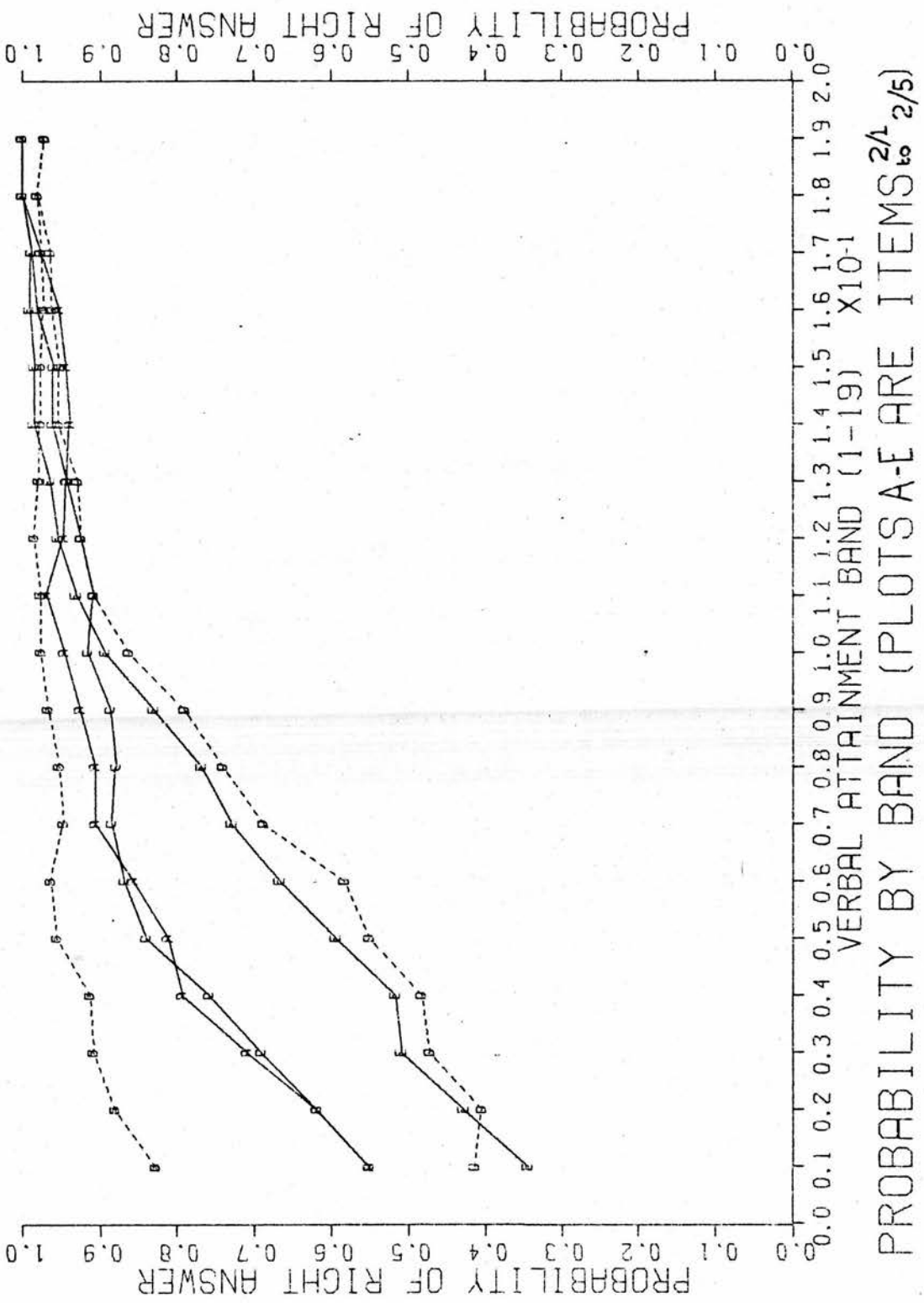

PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 2/1 to 2/5)
PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
VERBAL ATTAINMENT BAND (1-19) X10-1
PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 23.6

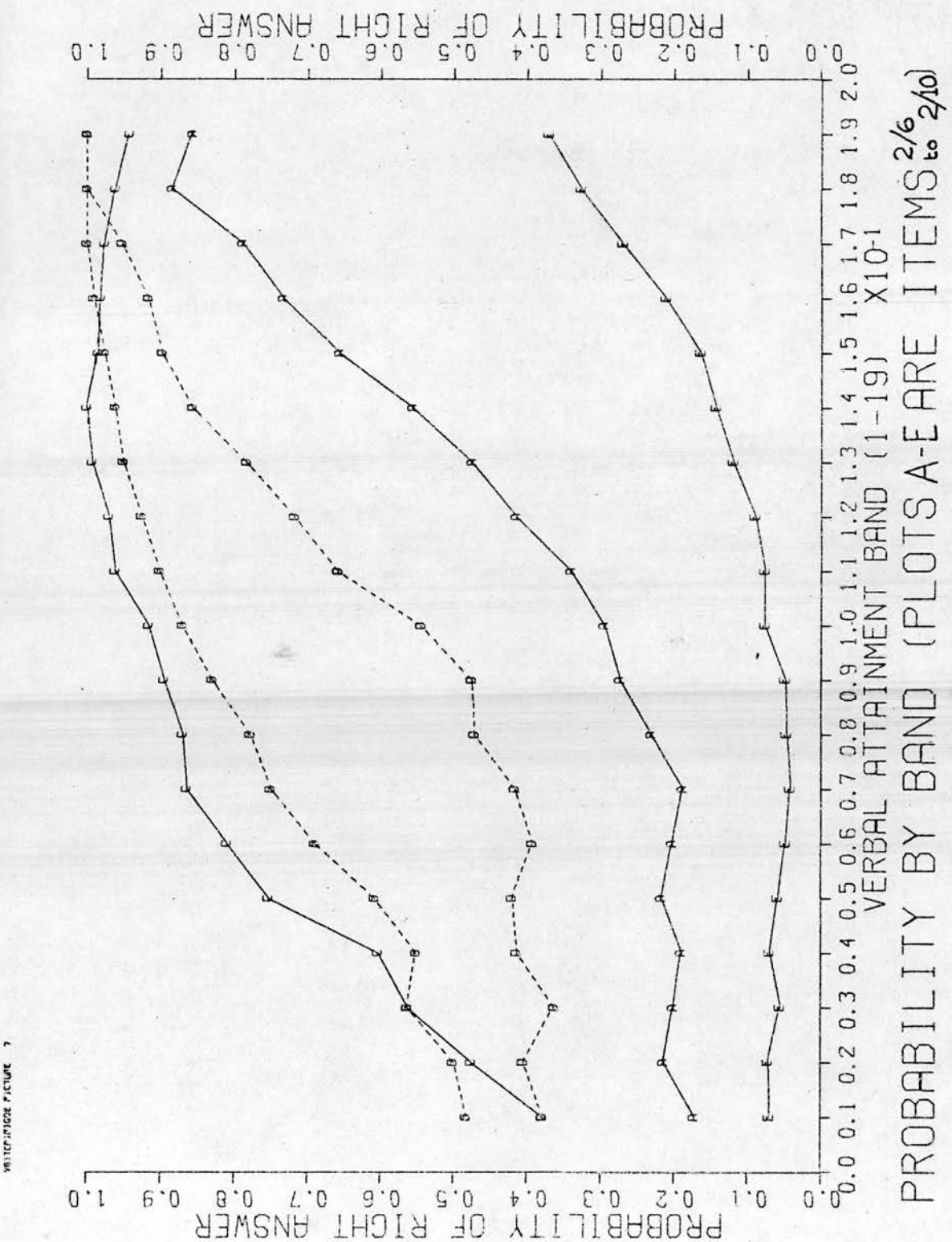

PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
VERBAL ATTAINMENT BAND (1-19) X10-1
PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 2/6 to 2/10)

FIGURE 23.7

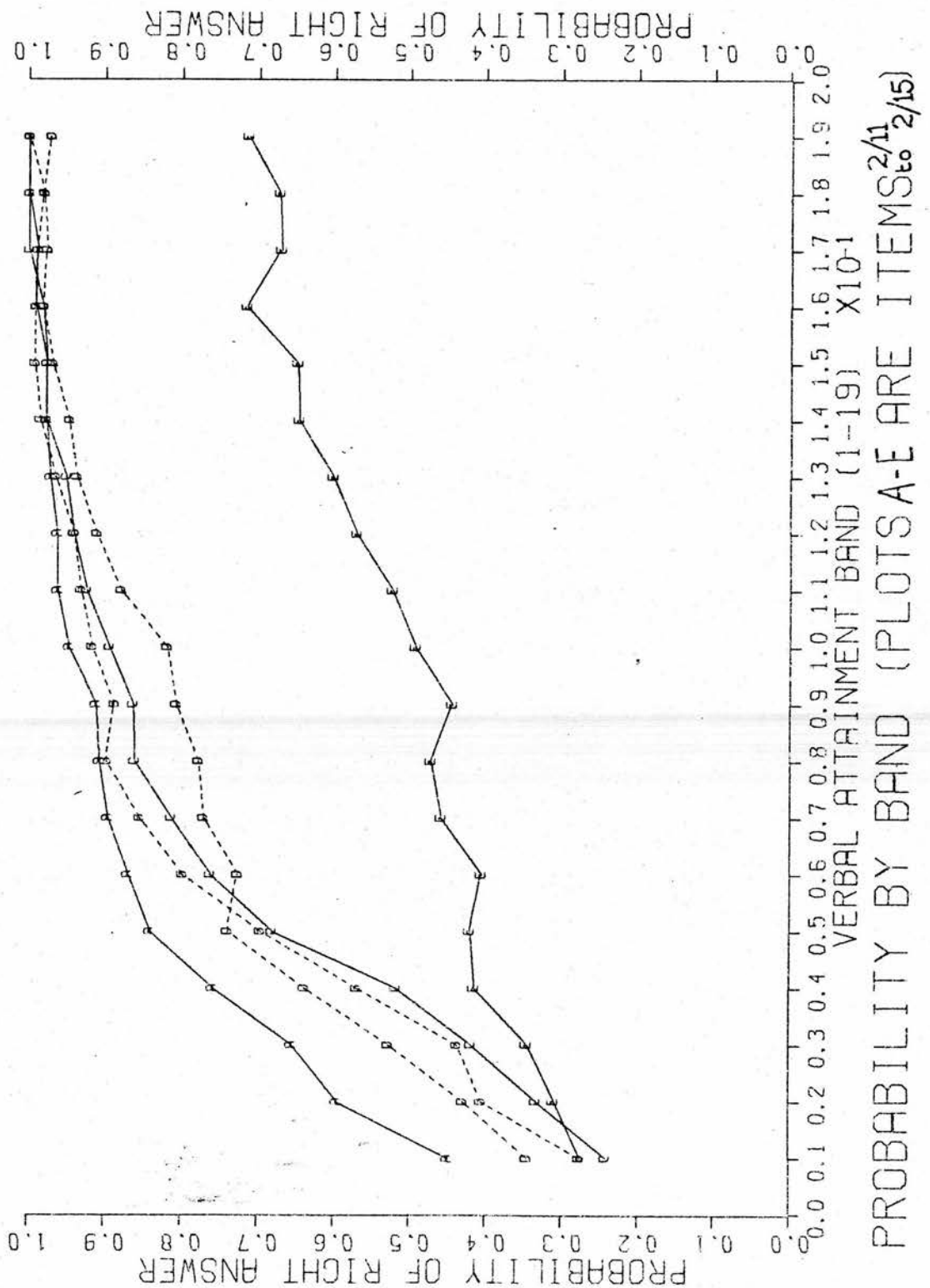
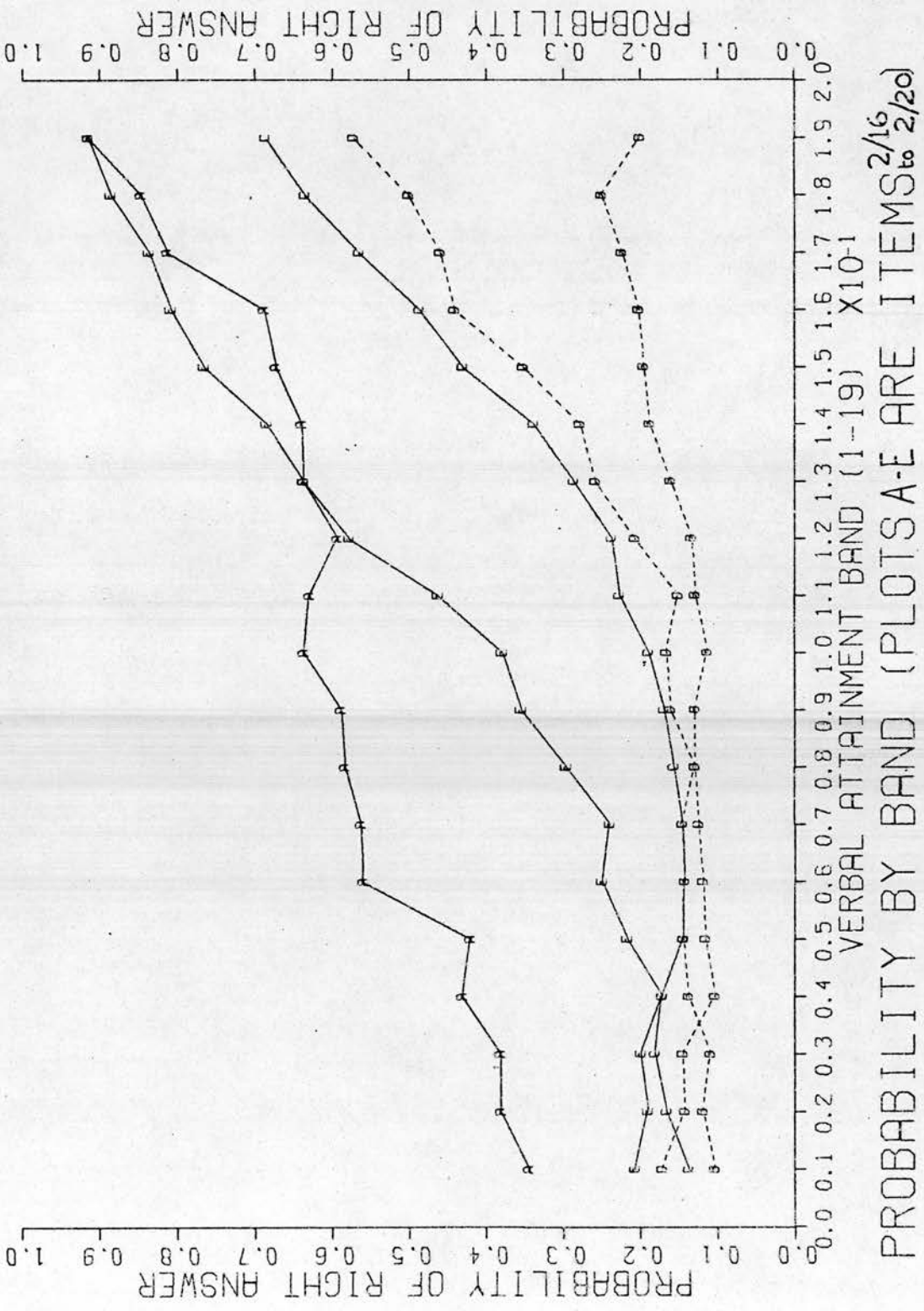

PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
VERBAL ATTAINMENT BAND (1-19) X10-1
PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 2/11 to 2/15)

FIGURE 23.8


PROBABILITY BY BAND (PLOTS A-E ARE ITEMS 2/16 to 2/20)
PROBABILITY OF RIGHT ANSWER
PROBABILITY OF RIGHT ANSWER
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
VERBAL ATTAINMENT BAND (1-19) X10-1

The Tail Location and Tail Discrimination indices depend for their evaluation on the specification of a percentile. Too extreme a percentile would make the indices too dependent on a small number of cases, and the resulting error of estimate would be high. Too central a percentile would lose the local tailishness of the index and tend too closely to the conventional global item statistics. To help decide on an appropriate percentile, tail indices were evaluated based on four different percentiles. The percentiles chosen were those that cut off a half, a quarter, an eighth, and a sixteenth of the derived distribution - numerically the 50th, 25th, 12·5th and 6·25th percentiles. The 50th percentile was included mainly for comparison. For convenience the attainment band values corresponding to these percentiles for an item/population derived distribution will be referred to as P2, P4, P8 and P16. These percentile values are indices of Tail Location: there are two sets for each item, one for the wrong-curve and one for the right-curve.

Given the choice of percentiles the index of Tail Discrimination can also be evaluated. The absolute value of the percentile/mean difference was taken as the index (as illustrated in Figure 17) and the four values of this index for the four chosen percentiles will be referred to as PMD2, PMD4, PMD8, and PMD16. Again there are two sets of values for each item.

PROGRAM 3 at Annex VIII takes in the conditional probabilities of item success on attainment and the overall population distribution of verbal attainment and, following the method described in Chapter 3 and above, produces the item/population derived distributions and the item tail characteristics.

Annex IX lists the proportionate wrong-curve and right-curve derived distributions (simple not cumulative) for the 240 items. Annex X gives the P-values (P2, P4, P8, & P16), and the PMD-values for the wrong-curve and right-curve for each of the 240 items.

The attainment band scale is, of course, discrete. The derived simple distributions (Annex IX) give the proportion falling at each of the 19 attainment bands. In determining P-values interpolation was carried out within a band in order to attribute an "exact" band-value to the location of the percentiles (50th, 25th, and so on). This interpolation assumed an even distribution within the band. This interpolated value will only be exact if the underlying continuous distribution is rectangular or stepwise - which it is not. However, the assumption of an underlying continuum is sound and the classification into 19 bands is considered fine enough to make the interpolation error negligible for the present purpose. In any case in the simulation the P-values will only be used to assign items to bands so that interpolation is not necessary for these: it is PMD-values for which the band unit is somewhat coarse, and the use of the interpolated values is to queue items in order within band. Minor inaccuracies in such queues are not considered important and fall within the worst-case philosophy.

In order to bring the tail characteristics (Annex X) into better perspective Tables 3 and 4 present frequency distributions of the P-values and the PMD-values.

Referring to Table 3 the R-distributions are less dispersed than the W-distributions. This reflects the occurrence of chance success which prevents the right-curve P-values reaching higher bands. The lower ends of the R-distributions and the upper ends of the

TABLE 3. Frequency distributions by attainment of the Tail Location of 240 items for selected percentiles

Attainment	P2[a]		P4		P8		P16	
Band[b]	W[c]	R[c]	W	R	W	R	W	R
1								
2								2
3	5							49
4	13		1			6		109
5	20		0			67		62
6	30		4	3	1	91		15
7	41		12	36	0	53		1
8	36		20	83	3	19		2
9	54		23	58	9	3	1	
10	38		38	41	18	1	2	
11	3	81	32	17	20		13	
12		68	46	2	31		15	
13		57	52		41		29	
14		31	12		43		33	
15		3			60		43	
16					14		75	
17							28	
18							1	
19								

Notes:

a P2, P4, P8 & P16 are the attainment band values at the 50th, 25th, 12·5th & 6·25th percentiles

b The nominal bands take in values down to 0·5 below and up to (but not including) 0·5 above the given band.

c The W and R columns refer to the wrong- and right-curves (as indicated in Figure 16).

TABLE 4. Frequency distributions by attainment of the Tail Discrimination of 240 items for selected percentiles. The index of Tail Discrimination used is the absolute value of the percentile/mean difference (PMD).

Attainment	PMD2[a]		PMD4		PMD8		PMD16	
Band[b]	W[c]	R[c]	W	R	W	R	W	R
0·0 - 0·5								
0·5 - 1·0					1		36	2
1·0 - 1·5					61	1	142	129
1·5 - 2·0			5		115	96	47	95
2·0 - 2·5	1		137	1	35	100	8	11
2·5 - 3·0	3		57	159	16	34	4	2
3·0 - 3·5	84	2	27	54	6	9	2	1
3·5 - 4·0	109	119	7	23	2		1	
4·0 - 4·5	23	93	3	3	3			
4·5 - 5·0	10	20	0		0			
5·0 - 5·5	6	5	2		0			
5·5 - 6·0	3	0	2		0			
6·0 - 6·5	0	0			0			
6·5 - 7·0	0	0			1			
7·0 - 7·5	1	1						
7·5 - 8·0								

Notes:

a PMD2, PMD4, PMD8 & PMD16 are the index values based on the 50th, 25th, 12·5th & 6·25th percentiles.

b The intervals are exclusive of the upper limit.

c The W and R columns refer to the wrong- and right-curves (as indicated in Figure 16).

W-distributions are the blunter. This effect is due to the limiting influence of the overall population distribution (see Figure 20) which contains all the tabulated distributions.

Referring to Table 4 it should be noted that it is the lower PMD-values that are of interest. These indicate blunter derived distributions and better tail-sweeping potential. The R-distributions tend to have higher PMD-values, and this represents the effect of chance success tapering the right-curves.

Tables 3 and 4 are helpful in formulating the selection criteria for picking the library items. Chapter 3.B.2 indicated that selection would be jointly on Tail Location and Tail Discrimination. Tables 3 and 4 give the individual rather than the joint distribution but they do indicate what values can be found. The aim would be to choose items with an even and wide-ranging spread of Tail Locations that also had average or better Tail Discrimination. Some compromise may be necessary in accommodating these joint aims, and clearly an even spread of Tail Location will not be fully achievable. There will be two separate item selections, one based on right-curve characteristics and the other on wrong-curve characteristics. Generally these selections would not be expected to contain the same items.

In picking the library items an equal number will be taken from each of the twelve tests. This is done so that the resulting attainment/response banks will represent all recruit samples equally. It is not likely that any sample would have a marked degree of eccentricity, but perhaps the same characteristics that would lead to over-representation of any test in the item library would be symptomatic of some fortuitously favourable set of circumstances. Unit weighting

of the test and recruit samples in the attainment/response banks to be used for the simulation appears to be a useful and readily available safeguard - although possibly an unnecessary one.

In fact four items will be selected from each test on the basis of their wrong-curve characteristics and four will be selected on the right-curve. Some overlap of items is possible. Given a little overlap this will mean a total item library of a little less than 96 items. This selection ratio is similar to that for constructing the 100-item test from the 240-item pool.

At the moment the method is comparing four alternative percentile bases for the evaluation of P-values and PMD-values. The next Chapter, Chapter 6: Results I, will present comparative analyses and decide on particular percentiles. The item library will then be selected on the corresponding indices according to the criteria outlined above.

C. Testing local independence in the item library

The assumption of local independence is that with ability held constant the probability of success on one test item is not predictable from success on another - the two item performances are independent. While not a necessary assumption for tailored testing it is nonetheless universally made because of its simplifying effect and its plausibility.

If the performance of N testees on two questions is as summarised in the following 2-by-2 frequency table,

	Item X		
Item Y	Wrong	Right	Total
Right	(WR)	(RR)	(RY)
Wrong	(WW)	(RW)	(WY)
Total:	(WX)	(RX)	N

then the probability of joint success on the two items may be estimated by (RR)/N. Given local independence the probability of joint success may also be estimated as the product of the two individual item success probabilities, that is by (RX)/N x (RY)/N. If local independence does not obtain then these two estimates will not be equatable. Statistical tests assuming the null hypothesis of independence of the dual classification can also be carried out to evaluate the probability of the observed or more extreme frequencies arising. Chi-square and Fisher exact probability tests would be appropriate for larger and small samples respectively.

The pairwise comparisons of item performance made here were necessarily made within tests - joint performance could only be looked at for items attempted by a common sample. The four items selected from each test for the item library (for their wrong-curve or right-curve tail characteristics) were considered in all six possible pairings. Attainment was held constant at each of six band levels, 3, 6, 9, 12, 15, and 18. At each band level both "narrow" band and "wide" band conditions were examined - a narrow band being the named band only, and a wide band including also the two neighbouring bands, that is a wide band attainment of 9, say, means attainment in bands 8 to 10.

Joint item performance data was produced for the six possible item pairings for each of the twelve tests, for each of the six chosen attainment bands, and for narrow and wide band conditions. This was done for both the wrong-curve set of items and the right-curve set. Altogether the joint performance of 144 item pairs (2 sets x 12 tests x 6 pairs) was looked at under twelve attainment conditions

(6 attainment levels x 2 band-widths). For each combination the four probabilities (RR)/N, (RX)/N x (RY)/N, (WW)/N, and (WX)/N x (WY)/N were evaluated and a Chi-Square value computed for the contingency table.

PROGRAM 4 at Annex XI was written to carry out this analysis working from the raw attainment/response data (Annex II) and information identifying the selected items.

The purpose of including the wide band attainment condition was to see if local independence would hold over the wide range. Once attainment is held constant the sample size for each attainment band is relatively small. Using a wide band triples the sample size and if local independence can be satisfactorily demonstrated for the less constant attainment of the wide band - with the benefit of more convenient statistical methods - then there will be no need for the less strenuous narrow band demonstration.

Although PROGRAM 4 computed a Chi-Square value for each case it was not always a meaningful statistic. Yates' continuity correction was included in the computation but nonetheless a minimum sample size of 40 is still desirable and even this will not be sufficient for very uneven cell frequencies in the contingency table. The Fisher exact probability test could be used with the smaller samples, say for samples of 30 or fewer, but less conveniently.

If, in the 500 or so wide band instances where the sample size will support a meaningful Chi-Square, the results support local independence then this will tend to be taken as sufficiently conclusive without need of further checking on the smaller samples. Chapter 7: Results II reports the findings of this aspect of the research.

D. Using response banks from recruit/library-item encounters for a real-data simulation of tailored testing

This Section describes the particular way in which the proposed tailored testing procedure was simulated using the data base. The aim is to specify an attainment level to be found and then to set the testing procedure to work finding it. The procedure will select appropriate questions and, in return, responses will be provided that were given by testees of the specified attainment. The procedure will update its derived distribution after each answer, and will terminate testing when this distribution has converged to the required precision. The interest is in how well the procedure converges on the attainment level specified. This level of attainment is inherent in the responses provided to the procedure and it is up to the procedure to identify the level satisfactorily and efficiently.

To allow evaluation of the procedure for testees of different attainment, three attainment levels were used. These were attainment bands 5, 10 and 15, chosen to represent below average, average and above average levels. Bands 5 and 15 are moderately extreme; 13% of recruits have attainments below 5, and a similar percentage has attainments above 15.

The chosen item library consists of two sets of 48 questions - the two sets being selected respectively for wrong-curve and right-curve characteristics. The two sets will be referred to as the W-set and the R-set. The item library classifies the questions: each set is arranged by attainment bands according to the Tail Location of the items, and within each band the items are queued in order of Tail Discrimination. Thus there are two ordered sets of items. The term

"set" is now further defined as referring to the ordered assembly of items.

By picking out the testees in attainment bands 5, 10 and 15 and noting their responses (simply right or wrong) to the item sets the required response banks can be assembled. Six response banks were set up. Three were for attainment bands 5, 10 and 15 - referred to as the narrow response banks for these bands: three were for bands 4 to 6, 9 to 11, and 14 to 16 - referred to as the wide response banks for bands 5, 10 and 15. The inclusion of the wide banks is to increase the size of the banks and so permit more simulated tests before a bank is exhausted. Of course, the attainment inherent in the wide bank responses will be less precise, but the interpretation of the results can take account of this.

Figures 24.1, 24.2 and 24.3 each illustrate a half of three of the response banks - Figure 24.1 is for an attainment level of narrow band 5 for the R-set items, Figure 24.2 is wide band 10 for the W-set, and Figure 24.3 is wide band 15 for the W-set. Generally the rows of the Figures correspond to individual items and the 0/1 entries in the row are wrong/right responses by different testees of the same specified attainment. That the item sets are ordered by Tail Location is reflected in the increasing proportion of wrong answers in the lower rows. That testees of different attainment levels are the donors of the responses is best seen by comparing Figures 24.2 and 24.3 which are both for the same item set. The response bank of Figure 24.1 holds about 1,000 responses, Figures 24.2 and 24.3 have about 3,500. The full response banks at these attainment levels are twice this size.

FIGURE 24 Three response banks. Rows correspond to items, entries 0/1 to wrong/right, 9 to empty.

FIGURE 24.1 Response bank for the R-set of items for attainment band 5 - narrow 5.

FIGURE 24.2 Response bank for the W-set of items for attainment bands 9 to 11 - wide 10.

FIGURE 24.3 Response bank for the W-set of items for attainment bands 14 to 16 - wide 15.

FIGURE 24.1

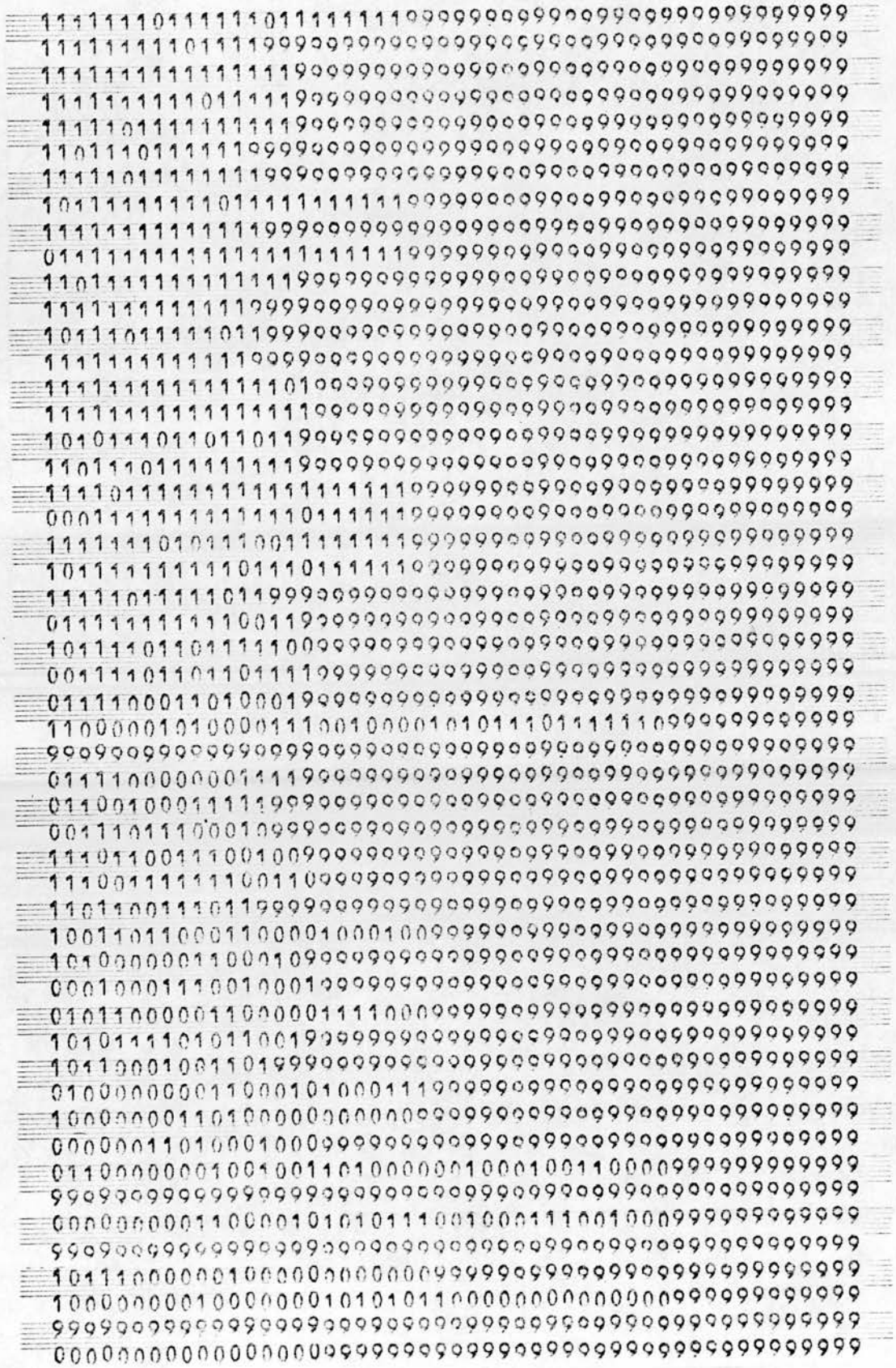
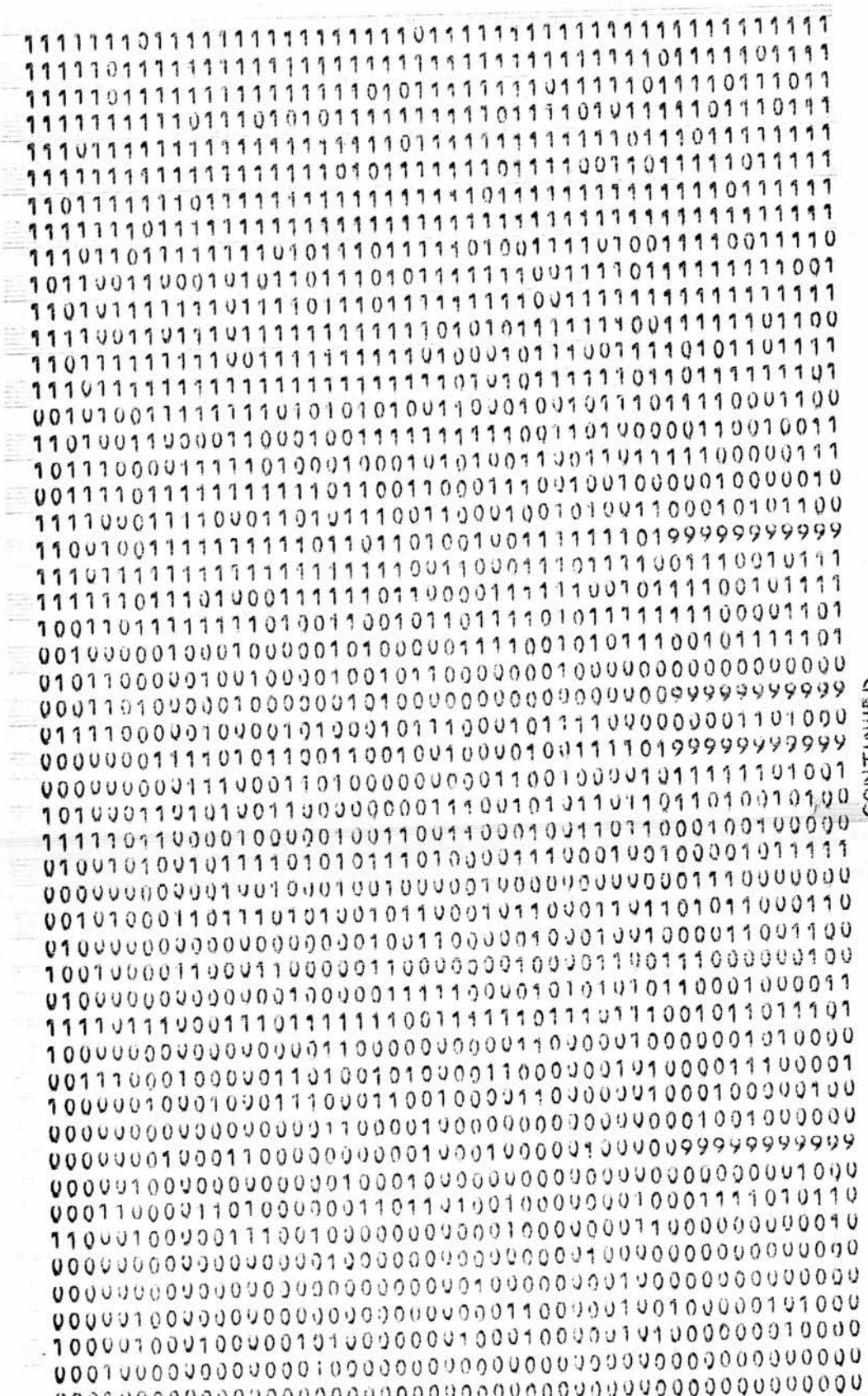
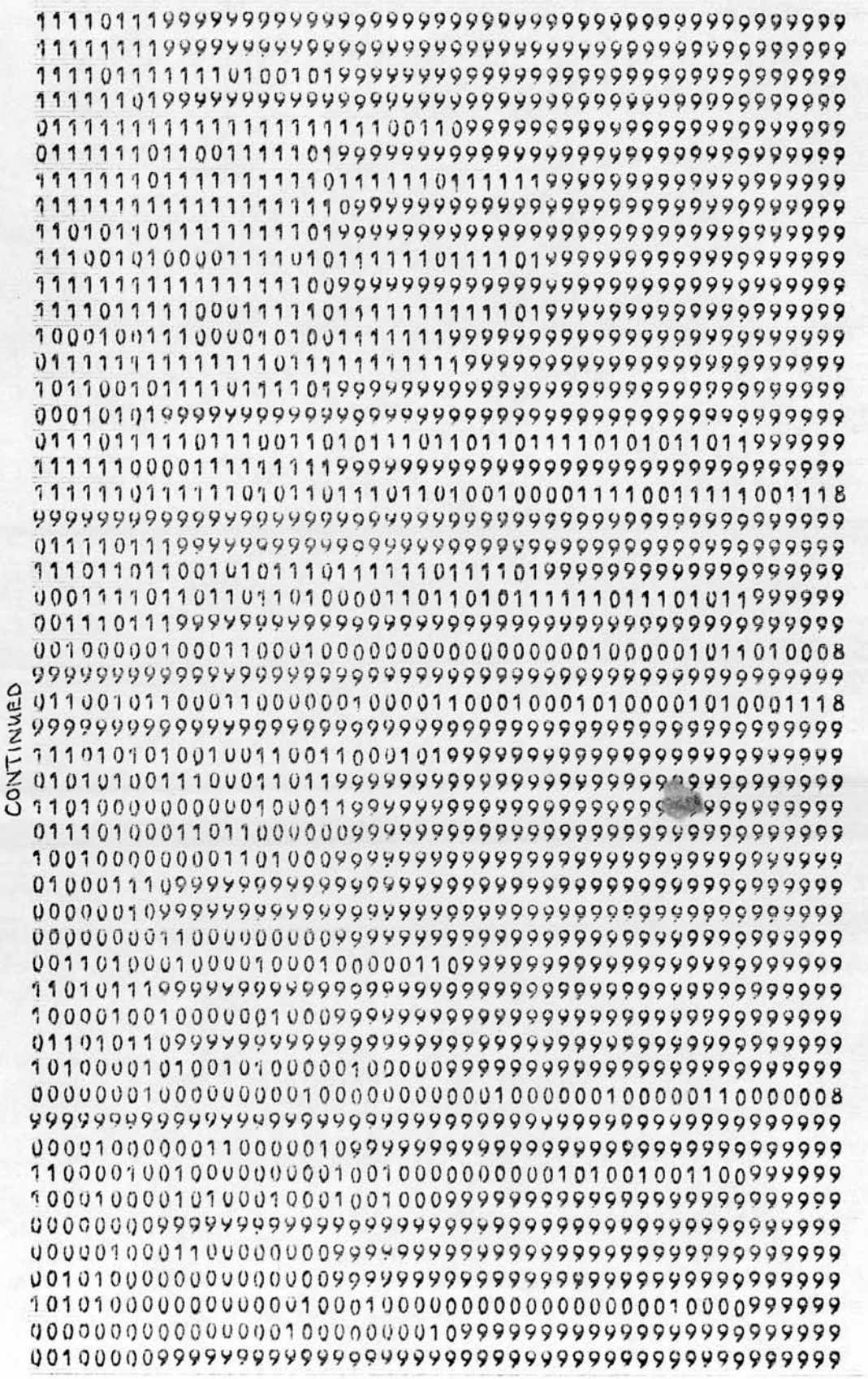


FIGURE 24.2



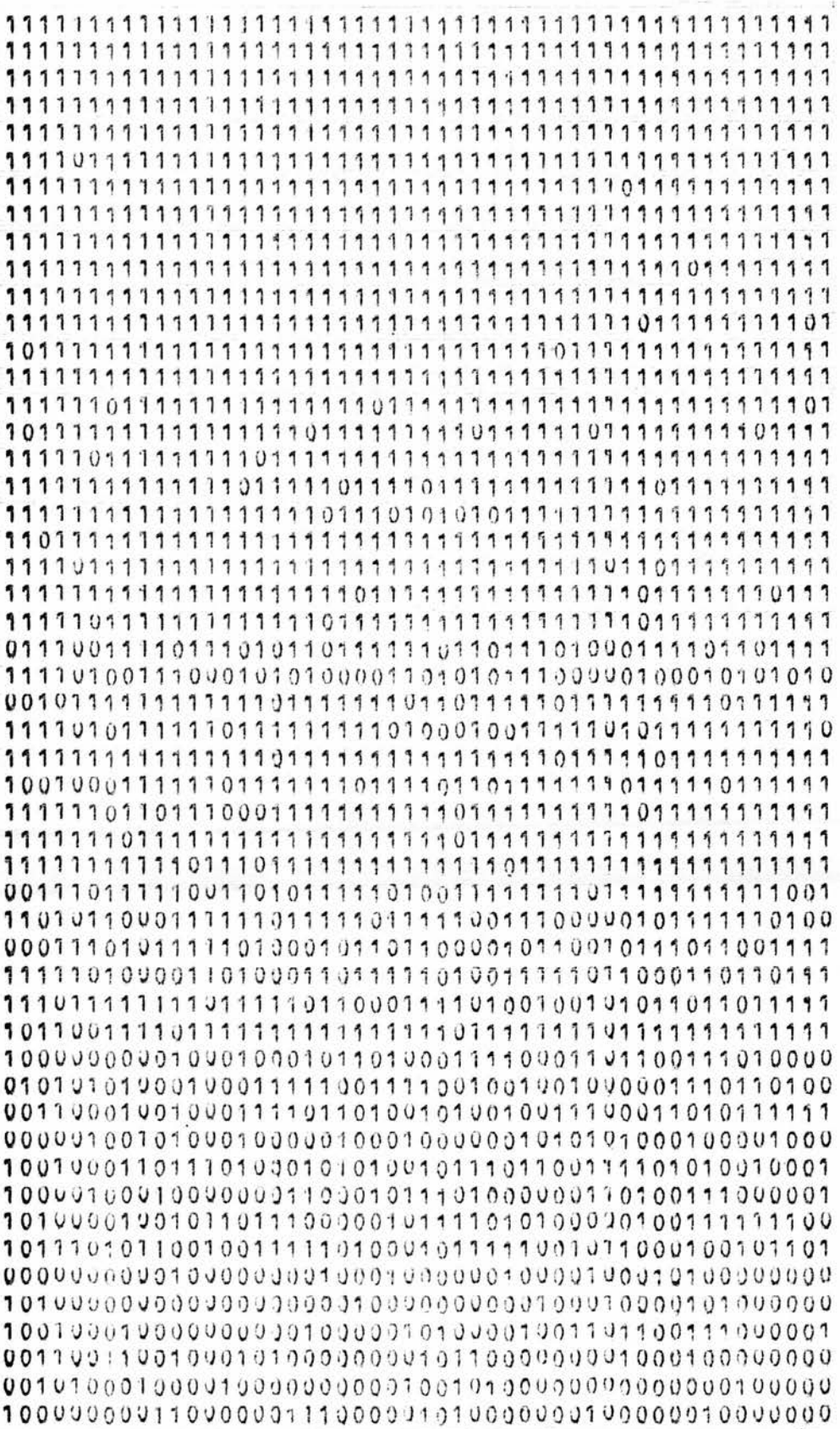
CONTINUED

FIGURE 24.2 continued.



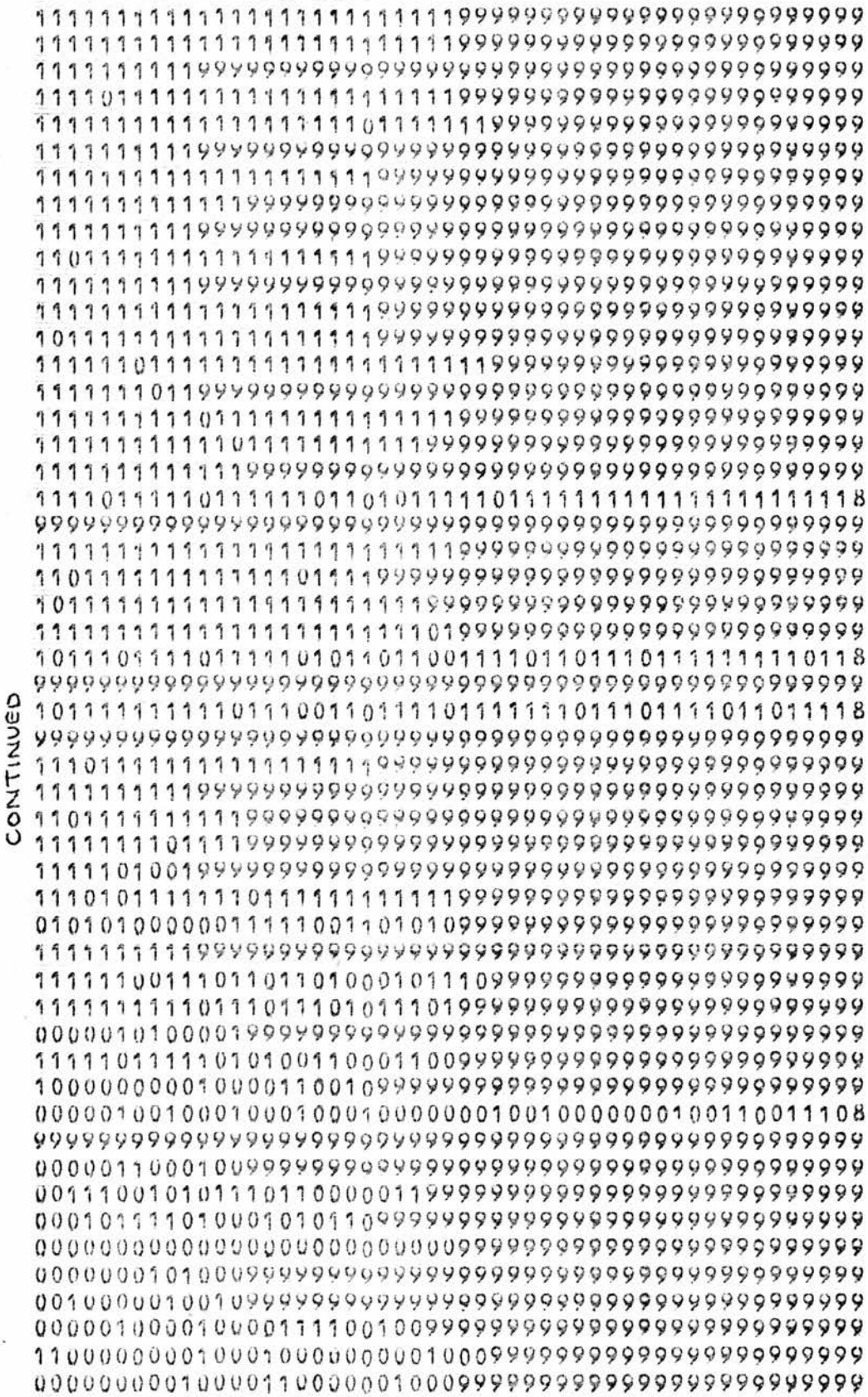
CONTINUED

FIGURE 24.3



CONTINUED

FIGURE 24.3 continued.



CONTINUED

For a specified attainment level the simulation calls on the corresponding response bank to provide right/wrong data for the successive questions chosen by the testing procedure. The response banks are sampled across testees, and for each attainment condition each response may be used once only. This makes each of the simulated tailored tests fully independent. Each test draws on a new portion of the response bank until the bank is exhausted. The size of the response bank determines how many such independent tests may be simulated.

Sampling responses across testees of the same measured attainment is a useful refinement compared with the usual practice of sampling within an individual testee's record. The point is that in live testing the true attainment level inheres in the testee and is sampled by a series of questions: the true attainment level is unknown but is by definition guaranteed in the series of responses because they are obtained at source. True attainment mediates the responses, and subject to some unreliability an estimate of true attainment may be obtained from them. However, in simulated testing the response bank perforce takes on the role of defining true attainment. Remember that in this study the attainment level associated with a recruit's responses is fully independent of those responses. We know that on retesting a proportion of the donors to a response bank would become ineligible because their reassessed attainment would be outside its ambit. By sampling across testees - who will display errors of measurement in both directions - the attainment level inherent in any response sample will be a better approximation to true attainment than may be the case within an individual testee record.

This sampling refinement basically tends to keep the simulated testing down to one error of measurement, as applies to live testing, rather than superimposing one such error on another.

Sampling across testees also allows the simulation to draw upon the same question several times within one test when the procedure repeatedly specifies that level of question. It is as if in live testing several questions were available with identical characteristics. This keeps nicely to the limitation of the kinds of questions that are likely to be available. A possible disadvantage of the approach is that some simulated tests may be overdependent on a relatively small sample of poor questions, but this is considered well compensated for by the increased flexibility of question use.

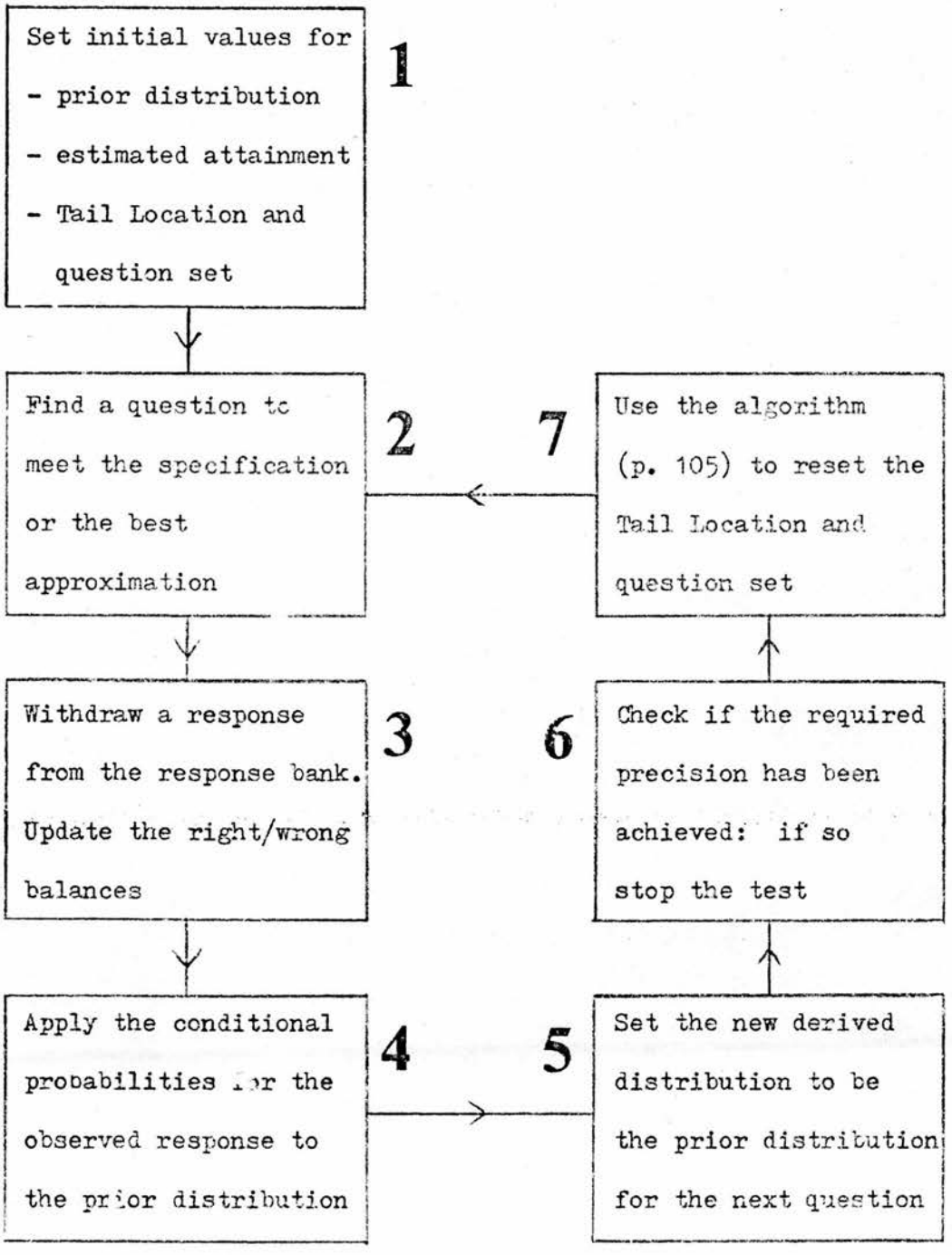
Figure 25 is a flow chart summarising the simulation of the proposed tailored testing procedure. This will help the further description of the method.

At stage 1 several quantities are given their starting values.

(i) The initial prior distribution is always set up as the overall population distribution portrayed in Figure 19.

(ii) The initial estimate of attainment is always set at one of the three bands 5, 10 or 15. However, in live testing, although an initial estimate will be available from the ACIO tests, this estimate will often be in error. It is necessary for the tailored testing procedure to be able to cope with erroneous estimates. To check this the estimates used were sometimes deliberate misestimates. There was thus a "true" attainment - defined by the response bank that would be used - and also an estimated attainment used in making a tailored start but subject to error. Table 5 gives the seven estimate /

FIGURE 25 Summary flow-chart for the tailored testing procedure.



misestimate attainment conditions used.

TABLE 5. Attainment conditions for the tailored testing simulations

Attainment level by band of the:-	
Response bank	Initial estimate
5	5
5	10
10	5
10	10
10	15
15	10
15	15

In addition to accurate estimates the 5 and 15 response bank levels are both subject to misestimates at average, band 10, level. The band 10 response bank is subject to both under- and over-estimates.

(iii) In order to select a question the test procedure uses two pieces of information, whether to look in the W-set or R-set of items, and in what attainment band the item Tail Location should fall. As indicated in Chapter 3 the R-set of items is used when a harder question is sought and vice-versa. The R-set is used to raise the lower bound of the ongoing attainment estimate. For the selection of the first question from the R-set the Tail Location is specified in relation to the initial estimate of attainment and its likely error. For verbal attainment the correlation between ACIO assessments and the standard verbal test scores obtained at selection centres is of the order of 0·8, however, such a high relationship does not occur for

other tests of the standard battery which have no direct ACIO counterpart. Consequently a lower correlation of only 0·6 is used as the basis for calculating error of estimate. On this basis a lower limit is calculated cutting off the same proportion of cases as the percentile chosen for the Tail Location index. (This calculation is the standard one in which the error of estimate is assumed normally distributed with a standard error of $SD\sqrt{1-r^2}$, where SD is the standard deviation of the population distribution.) In anticipation of the findings of the next chapter, Results I, this initial lower limit is set 4 bands below the initial estimate of attainment. The initial upper limit is similarly set 4 bands above the estimate and provides the location for selecting the first question from the W-set. In starting the test procedure it is initially directed to look for a question in the R-set with its Tail Location at the band given by the lower limit calculated above.

Thus for the misestimate conditions the initial limits do not in fact include the true value. For example, for a true attainment of band 10 with an initial misestimate of 5, the initial lower and upper limits will be 1 and 9. This should provide a rigorous test of the procedure.

At stage 2 when a question is sought with a particular Tail Location there will sometimes be no exact match. This might happen in live testing simply because there are no questions there or because the available questions have all been tried, and it happens in simulated testing for similar reasons - either deficiences in the item library or a sector of the response bank has been exhausted. The testing procedure has to be able to cope with this eventuality - otherwise it does not meet its specification of not having critical

item requirements - and does so by finding a question at the closest approximation to the band specified. This approximation is defined as follows:-

i. unless two alternative band locations are equally near take the nearest,

and where they are equally near,

ii. if a change of Tail Location is being made take the question further in the direction of the change,

iii. and if no change is being made choose the easier.

At stage 3 a response to the located question is withdrawn from the response bank. The balance of right and wrong answers is now updated, both overall and at the particular band level. This will provide the data for deciding the specification for the next question at stage 7.

The simulation has also to update its own records of questions and responses used to ensure that no responses are used twice, and to expedite the next question choice.

Stages 4 and 5 are straightforward applications of what was described in Chapter 3 and require no further comment.

At stage 6 a decision is made as to whether or not a test can stop. Testing stops when a required precision is achieved. The required precision is that of the 100-item conventional standard verbal test described earlier. This had a parallel forms reliability estimated at 0·94 (p. 113). The parallel forms reliability seems an appropriate form on which to base the precision the tailored test should match because, in live tailored testing, it will generally be that even a retest of the same testee will be based on a possibly

overlapping but different sample of items. It is the item sampling aspect of a tailored test that indicates a form of reliability taking adequacy of item sampling into account. In any case it is true for all but highly specific tests, conventional or otherwise, that they aspire to be independent of the particular item sample they embrace.

The standard error of measurement of the 100-item standard verbal test is thus 1·12 in band units (that is, $SD/\sqrt{1-r}$). This gives a 90% confidence interval 3·8 bands wide. The test simulations will thus be stopped when the 5th and 95th percentiles of the derived distribution have converged to a separation of 3·8 or less. Generally the simulated tests will necessarily stop with a fractionally greater precision than that of the 3·8 value.

Stage 7 employs the algorithm of chapter 3 (p. 105) on the right/wrong balances from stage 3 to determine the item set and Tail Location for the next question. When Tail Location has to be changed this is done in units of two bands. Some evidence for this choice is given later, but no claim is made that it is optimum only that it is satisfactory. Some compromise is required that can both rectify initial misestimates quickly and can also settle to a fairly even level of question difficulty. The item set and Tail Location information enters at stage 2 and the testing cycle repeats until the termination criterion is met.

Two further glosses are required on the method to complete the description. Because of the way they were chosen the R-set of items is somewhat easier than the W-set. In part this represents a deficiency in the item library, but again the test procedure is required to cope with such deficiences. The device incorporated in the tailoring

process to deal with this imbalance is to permit a change of item set. If the initial rule of seeking harder items, say, within the R-set cannot bring the overall number of right and wrong answers into balance then the procedure is allowed to change over to the W-set. This is referred to as a tail change and can happen only once during a test. Testees of well above average ability are likely to change over to the W-set regularly and testees of well below average to change over to the R-set. The procedure makes a tail change when the imbalance between right and wrong answers reaches nine either way. This is set at such a relatively high trigger value to avoid inappropriate tail changes. When stage 3 updates the right/wrong balances a check is also made to see if a tail change is called for.

The second addition was made only for the simulation and is not relevant to live testing. In the simulation it was desired to check when the response banks were becoming seriously depleted. It was visualised that when a response bank had serviced some number of tests the matching of the questions sought to the questions for which responses remained would become increasingly approximate. In this situation it was determined to stop testing if a right/wrong imbalance as large as 18 arose. This then was a secondary termination criterion used to detect an impoverished response bank.

The computer program written to implement the details of this section is PROGRAM 5 and appears at Annex XII. The results of the simulations are presented and discussed in Chapter 8: Results III.

6. RESULTS I AND DISCUSSION: SELECTING THE ITEM LIBRARY, AND A COMPARISON OF CONVENTIONAL ITEM CHARACTERISTICS AND ITEM TAIL CHARACTERISTICS.

The two item tail characteristics, Tail Location and Tail Discrimination, were introduced earlier. The quantitative indices being used for these characteristics are a chosen percentile and an absolute percentile/mean difference (PMD), (Chapter 3. B.2). The final definition of the indices requires a specific percentage value to be prescribed for each index. Four possible values have so far been calculated in parallel: these values are based on the 50th, 25th, 12·5th and 6·25th percentiles. The nomenclature adopted (Chapter 5.B) is that the Tail Location indices for these percentiles are referred to as the P-values, P2, P4, P8 and P16, and the Tail Discrimination indices as the PMD-values, PMD2, PMD4, PMD8 and PMD16.

A. Specifying the tail indices and a first comparison with conventional item characteristics

The first part of this chapter compares the four possibilities for each index and decides on one. The general principle governing the choice is to find values which are good representatives of their class. The two indices should also have some substantial measure of independence from each other and from conventional global item characteristics.

The relationships between the two tail indices, the P-values and the PMD-values, and two conventional item characteristics were examined for all 240 items of the item pool. The two conventional characteristics were item easiness (indexed by the percentage of recruits answering an item correctly), and item discrimination (indexed by the

point-biserial correlation between item performance and the standard verbal test score). These two global characteristics will be referred to as overall easiness and overall discrimination - or simply easiness and discrimination. The values for overall item easiness and discrimination were available from a routine item analysis of the 240 items - except that some minor adjustments were made to the easiness values[1].

Figures 26.1 to 26.40 plot the relationships among the tail indices and conventional characteristics. All pairings are considered other than that between overall easiness and discrimination. Within each of the five pairings the P-values and/or PMD-values have the four possible values being compared: there are also the two sets of the P- and PMD-values, one for the wrong-curve and one for the right-curve. Thus there are forty plots (5 pairings x 4 values x 2 sets). The plots are fully identified in the general description at the start of Figure 26.

Some of the comments previously made on the tail indices in relation to Tables 3 and 4 are reflected in the plots - for example, the smaller dispersion of the right-curves resulting from the possibility of chance success, and also the bunching of low PMD-values at low right-curve bands and high wrong-curve bands caused by the constraint of the enveloping population distribution. Generally the limiting influence of the population distribution is apparent in many plots,

---

1. The recruit samples obtained during pretesting (pp. 109-110) matched approximately but not exactly the population verbal attainment distribution taken as definitive (Figure 19). Adjustments were made to the sample easiness values to allow for this, but these were small and while strictly accurate are considered of no practical significance here.

FIGURE 26

Scatterplots of the relationships between tail indices and conventional item characteristics. (Tail Location indices are denoted Tail Percentiles.)

	W-set Figures	R-set Figures
Tail Discrimination x Tail Percentiles	26•1 - 26•4	26•21 - 26•24
Overall easiness x Tail Percentiles	26•5 - 26•8	26•25 - 26•28
Overall discrimination x Tail Percentiles	26•9 - 26•12	26•29 - 26•32
Tail Discrimination x Overall easiness	26•13 - 26•16	26•33 - 26•36
Tail Discrimination x Overall Discrimination	26•17 - 26•20	26•37 - 26•40

Within each set of four graphs the tail indices are based in turn on the 50th, 25th, 12•5th and 6•25th percentiles. This is indicated on the graphs by P2, P4, P8 and P16. Reference to the W-set or R-set is indicated by W: or R: prefixing the P-value. The values of Pearson r are given on each graph.

Figures 26.1 to 26.40 follow:-

FIGURE 26.1

W:P2
r=-.23
RELATION BETWEEN TD INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX X10-1
TAIL PERCENTILES X10-1
0.24
0.28
0.32
0.36
0.40
0.44
0.48
0.52
0.56
0.60
0.64
0.68
0.72
0.76
0.80
0.84
0.88
0.92
0.96
1.00
1.04
.250
.300
.350
.400
.450
.500
.550
.600
.650
.700
.750
.800
.850
.900
.951
1.001
1.051
1.101
1.151
1.201
1.25

FIGURE 26.2

W:P 4
r=-.54
RELATION BETWEEN TD INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0
6.4
6.8
7.2
7.6
8.0
8.4
8.8
9.2
9.6
TAIL PERCENTILES X10-1
.00
40.
80.
120.
160.
200.
240.
281.
321.
361.
401.
441.
481.
521.
561.
601.
641.
681.
721.
761.
801.
841.
881.
922.00

FIGURE 26.3

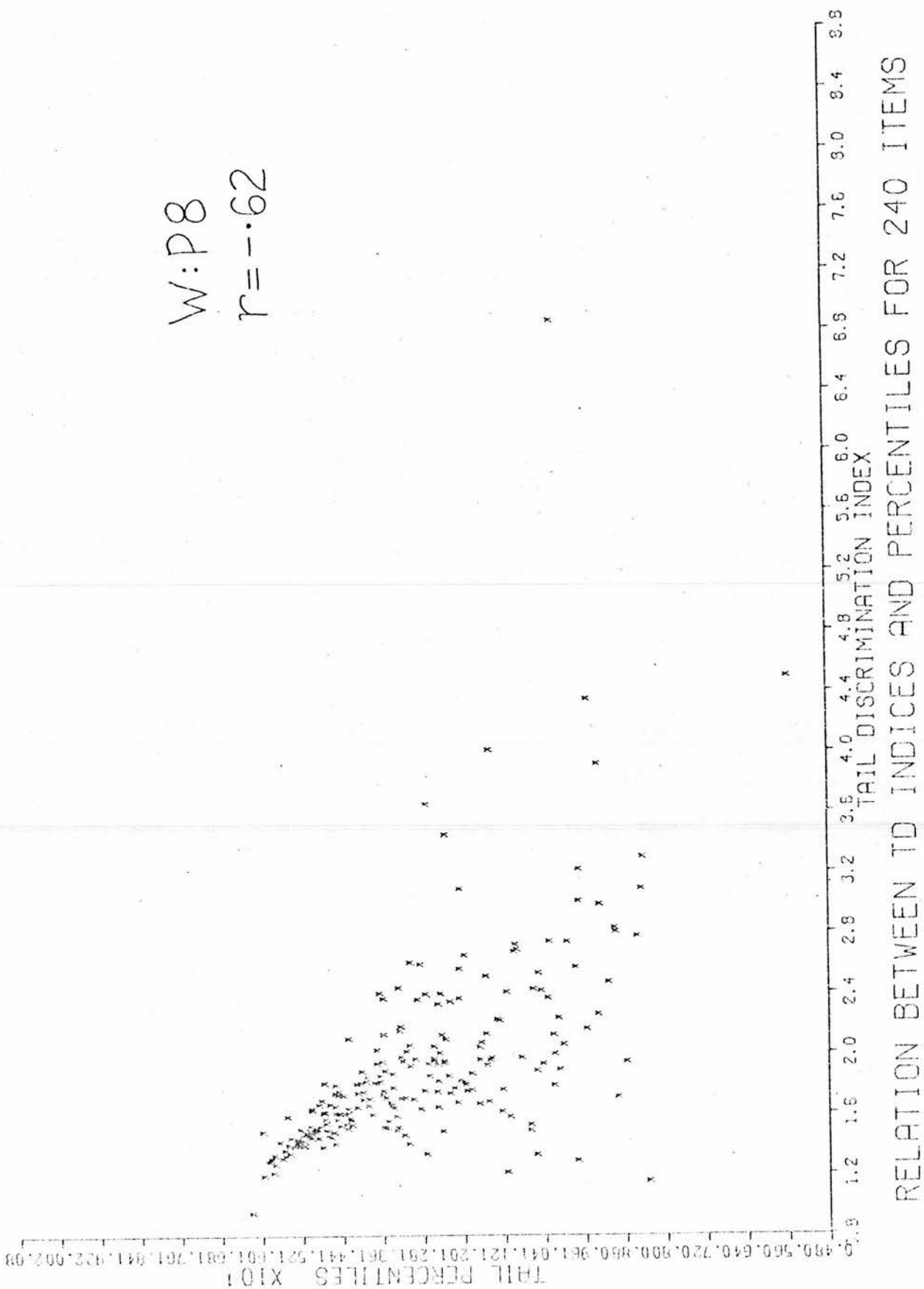

W:P8
r=-.62
RELATION BETWEEN TD INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0 8.4 8.8
TAIL PERCENTILES X10¹
.9 .480 .560 .640 .720 .800 .880 .961 1.041 1.121 1.201 1.281 1.361 1.441 1.521 1.601 1.681 1.761 1.841 1.922 2.002 2.08

FIGURE 26.4

W:P16
r=-.67
RELATION BETWEEN TD INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
TAIL PERCENTILES X10-1
.99
.951
.901
.851
.801
.751
.701
.651
.601
.551
.501
.451
.401
.351
.301
.251
.201
.151
.101
.051
.001
.000

FIGURE 26.5

RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10-1
.250 .300 .350 .400 .450 .500 .550 .600 .650 .700 .750 .800 .850 .900 .951 1.001 1.051 1.101 1.151 1.201 1.25
W:P2
r=-.84

FIGURE 26.6

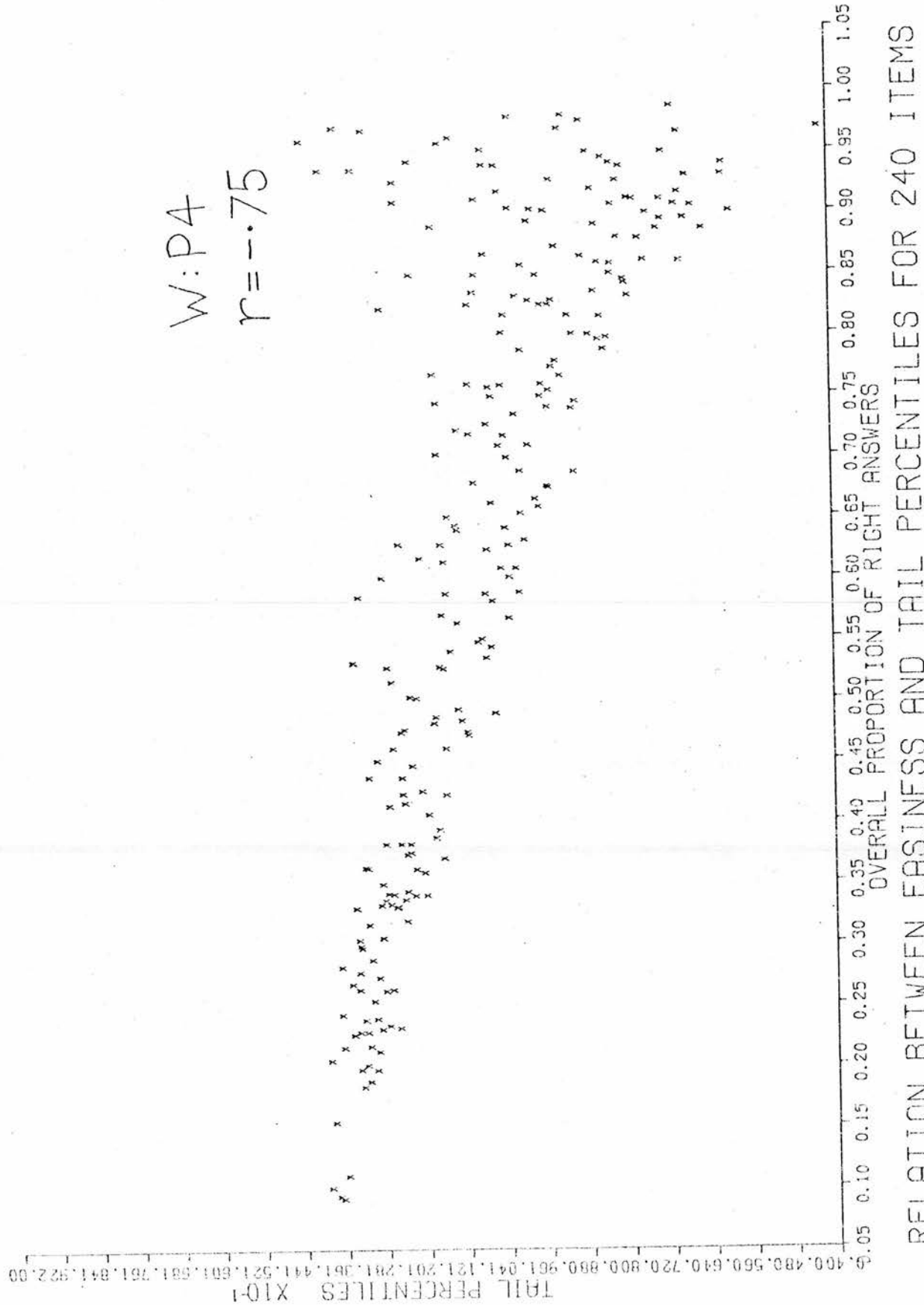

RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
W:P4
r=-.75
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10-1
.400 .480 .560 .640 .720 .800 .880 .961 .041 .121 .201 .281 .361 .441 .521 .601 .681 .761 .841 .922 .00

FIGURE 26.7

W:P8
r=-.67
RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10³
.08 .002 .922 .841 .761 .681 .601 .521 .441 .361 .281 .201 .121 .041 .961 .880 .800 .720 .640 .560 .480

FIGURE 26.8

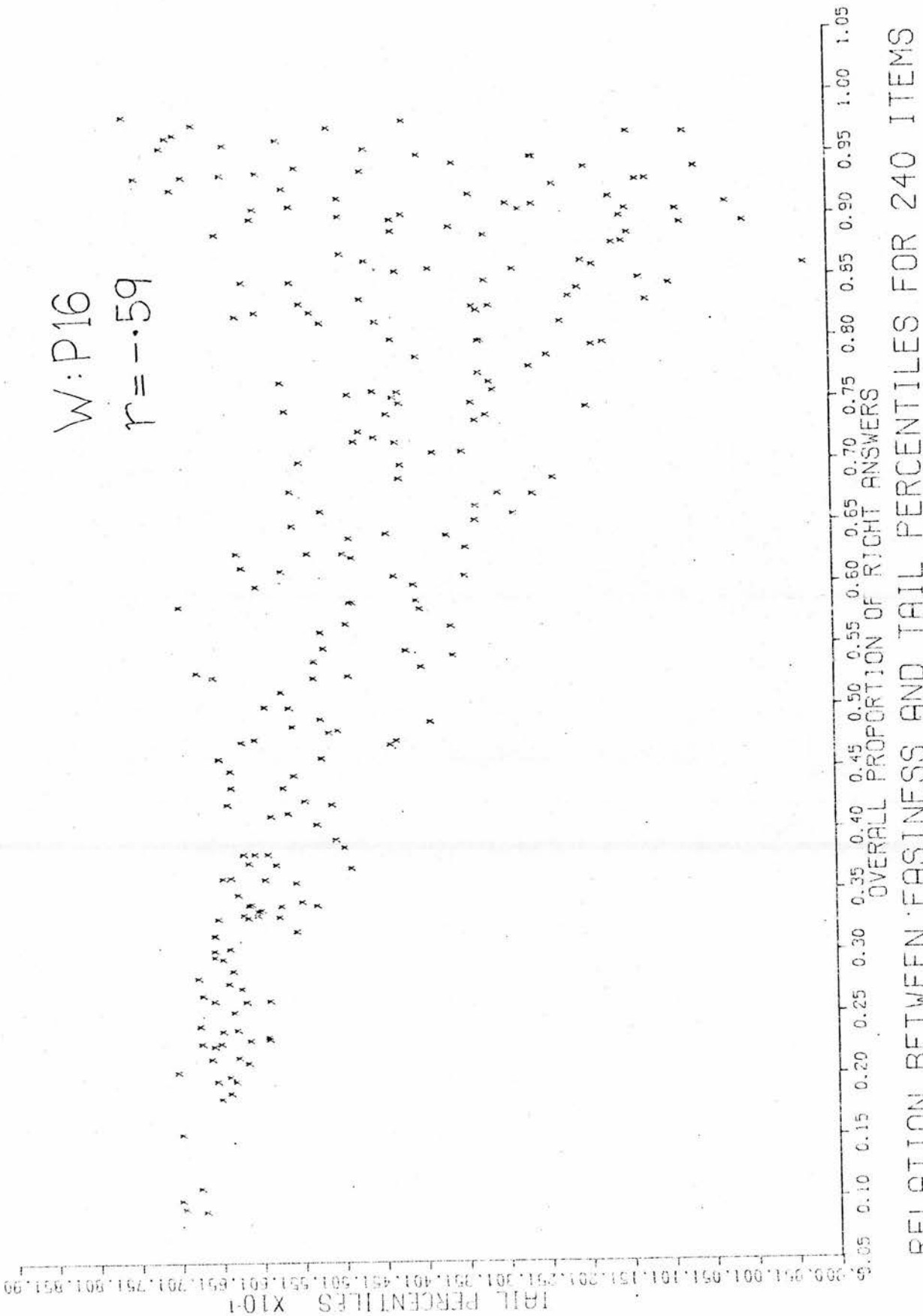

W:P16
r=-.59
RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10-1
.99 .95 .90 .85 .80 .75 .70 .65 .60 .55 .50 .45 .40 .35 .30 .25 .20 .15 .10 .05

FIGURE 26. 9

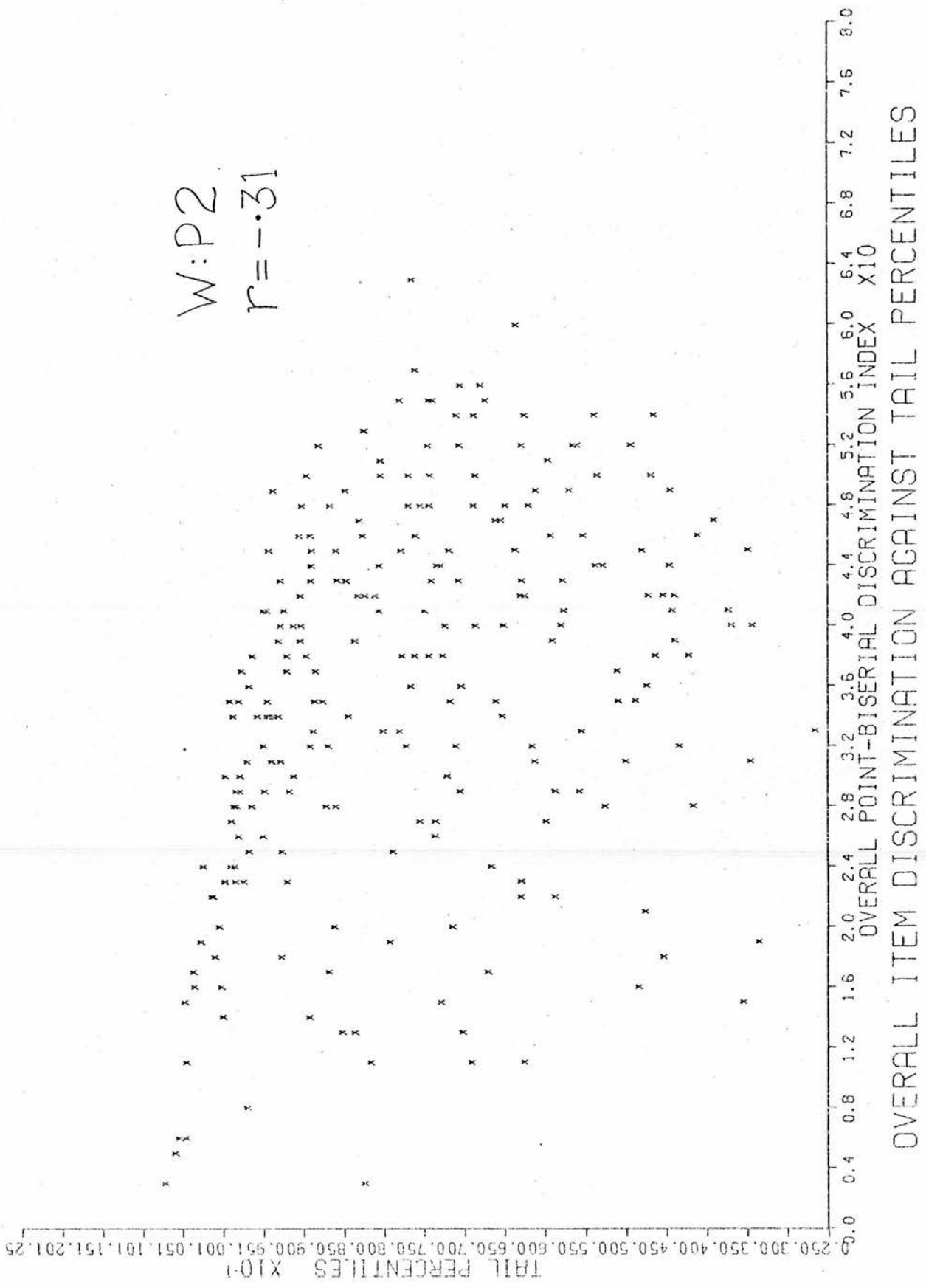

W:P2
r=-.31
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
TAIL PERCENTILES X10-1
0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0
250 300 350 400 450 500 550 600 650 700 750 800 850 900 951 001 051 101 151 201 25

FIGURE 26.10

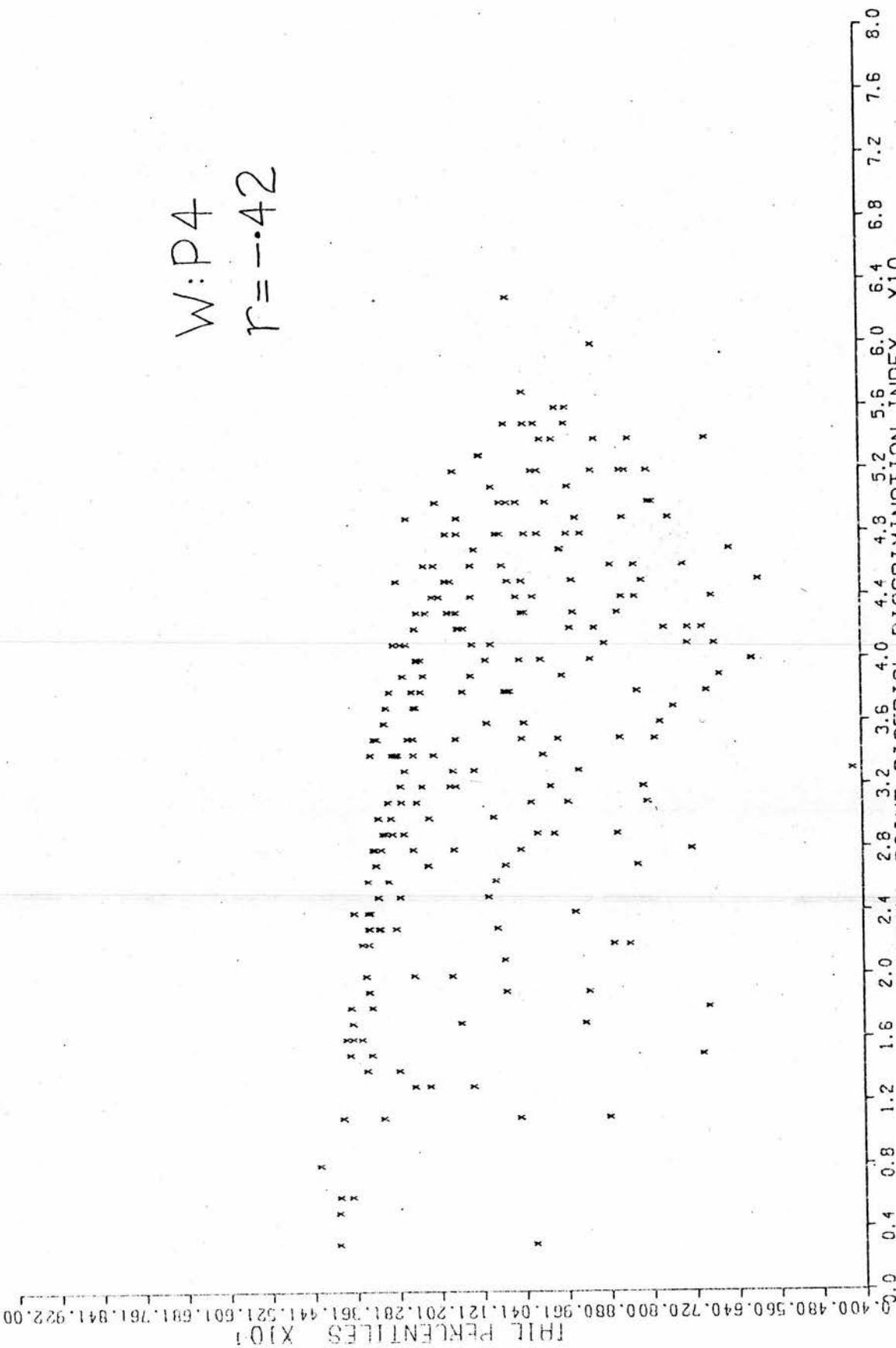

W:P4
r=-.42
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0
TAIL PERCENTILES X10
0.00 40.80 161.41 281.61 401.81 521.41 641.61 761.41 880.80 1000.40

FIGURE 26.11

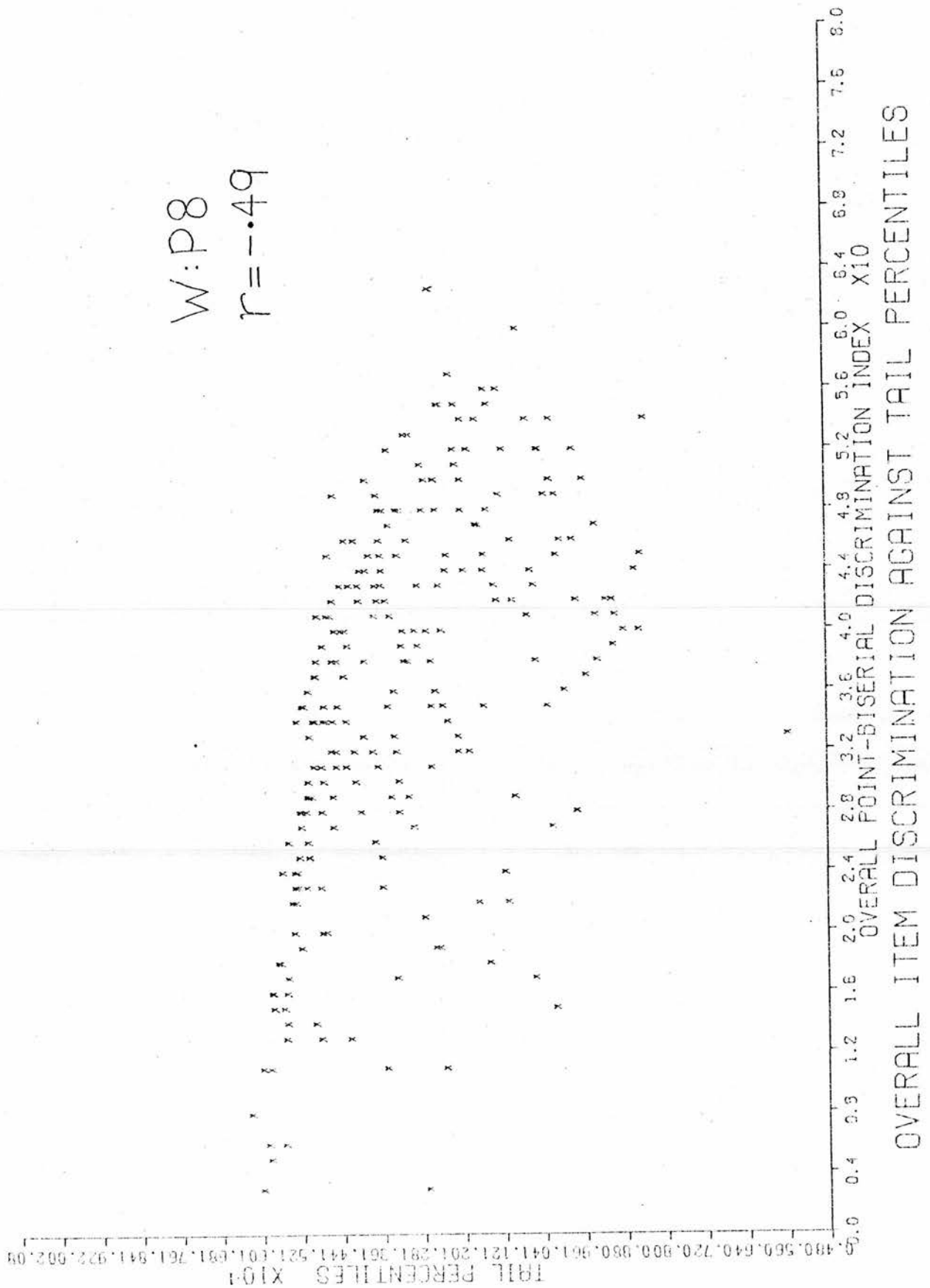

W:P8
r=-.49
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
TAIL PERCENTILES X10-1
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0
6.4
6.8
7.2
7.6
8.0
.480,560,640,720,800,880,961,041,121,201,281,361,441,521,601,681,761,841,922,002,08

FIGURE 26.12

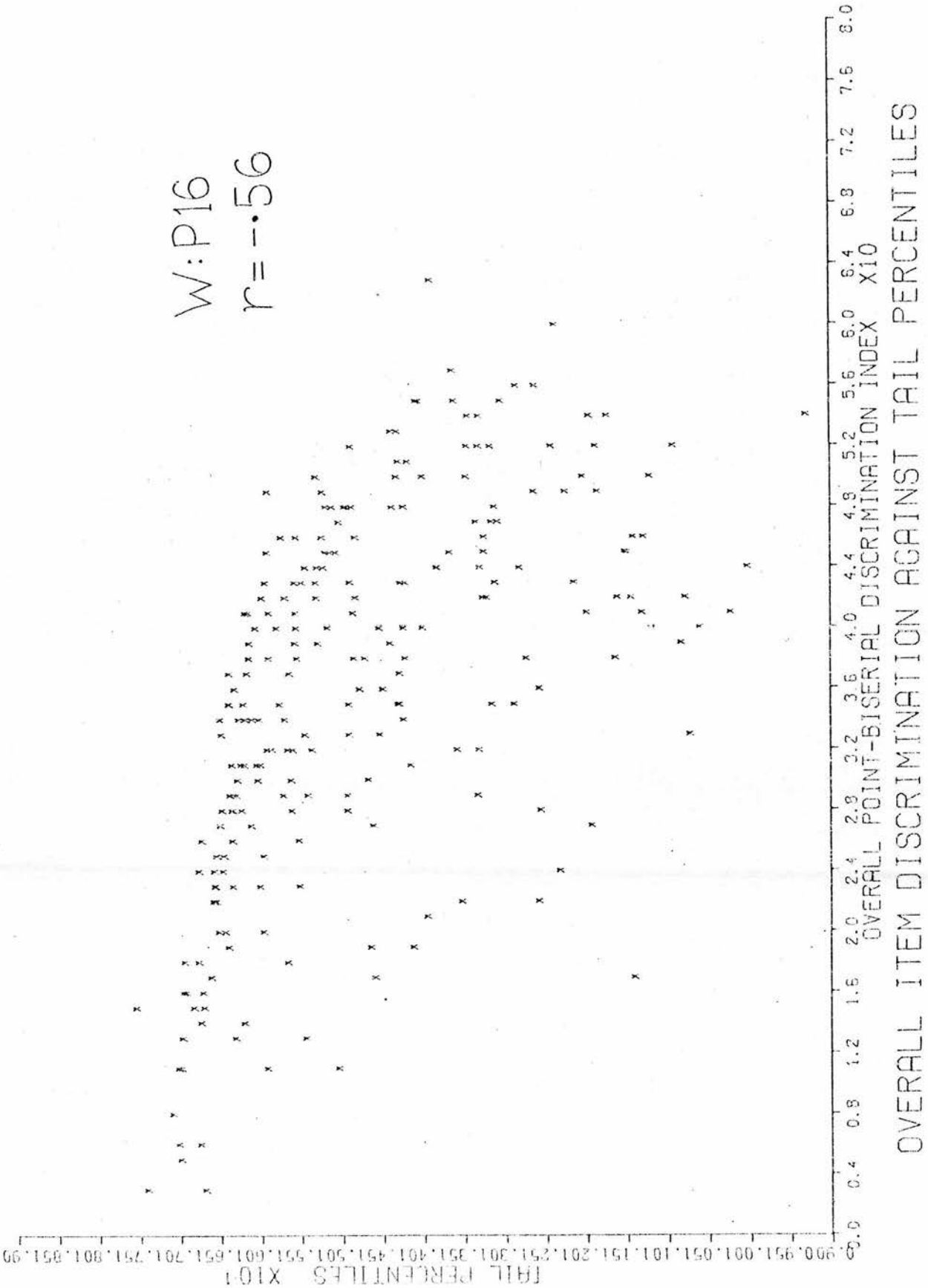

W:P16
r=-.56
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
TAIL PERCENTILES X10
0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0
.90 .95 .00 .05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60 .65 .70 .75 .80 .85 .90

FIGURE 26.13

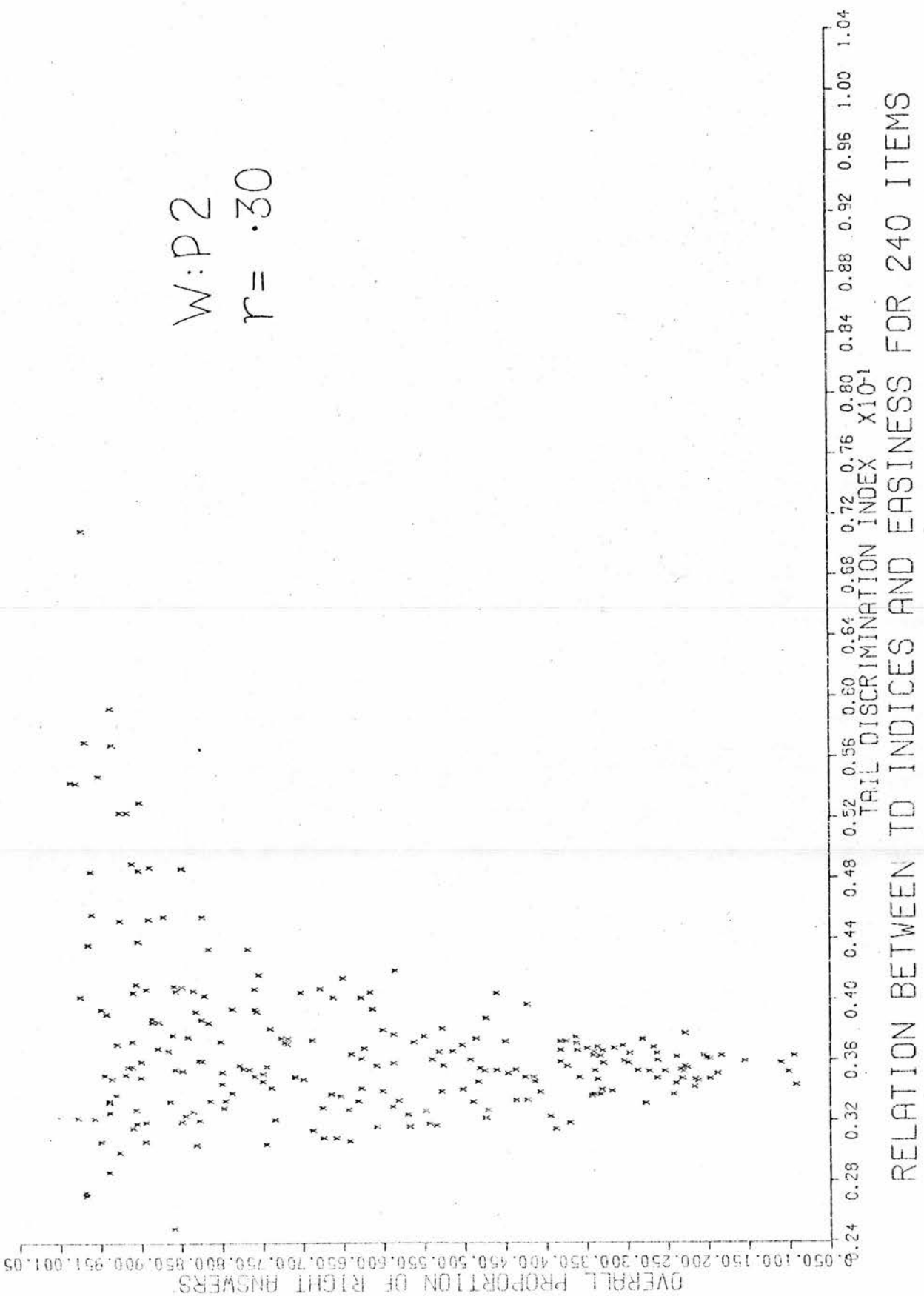

W:P2
r= .30
RELATION BETWEEN TD INDICES AND EASINESS FOR 240 ITEMS
TAIL DISCRIMINATION INDEX X10-1
0.24 0.28 0.32 0.36 0.40 0.44 0.48 0.52 0.56 0.60 0.64 0.68 0.72 0.76 0.80 0.84 0.88 0.92 0.96 1.00 1.04
OVERALL PROPORTION OF RIGHT ANSWERS
.050 .100 .150 .200 .250 .300 .350 .400 .450 .500 .550 .600 .650 .700 .750 .800 .850 .900 .951 1.001.05

FIGURE 26.14

RELATION BETWEEN TD INDICES AND EASINESS FOR 240 ITEMS
W:P4
r= .54
TAIL DISCRIMINATION INDEX
OVERALL PROPORTION OF RIGHT ANSWERS
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0
6.4
6.8
7.2
7.6
8.0
8.4
8.6
9.2
9.6
.05
.001
.951
.900
.850
.800
.750
.700
.650
.600
.550
.500
.450
.400
.350
.300
.250
.200
.150
.100
.050
0

FIGURE 26.15

W:P8
r = .51
RELATION BETWEEN TD INDICES AND EASINESS FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
OVERALL PROPORTION OF RIGHT ANSWERS

FIGURE 26.16

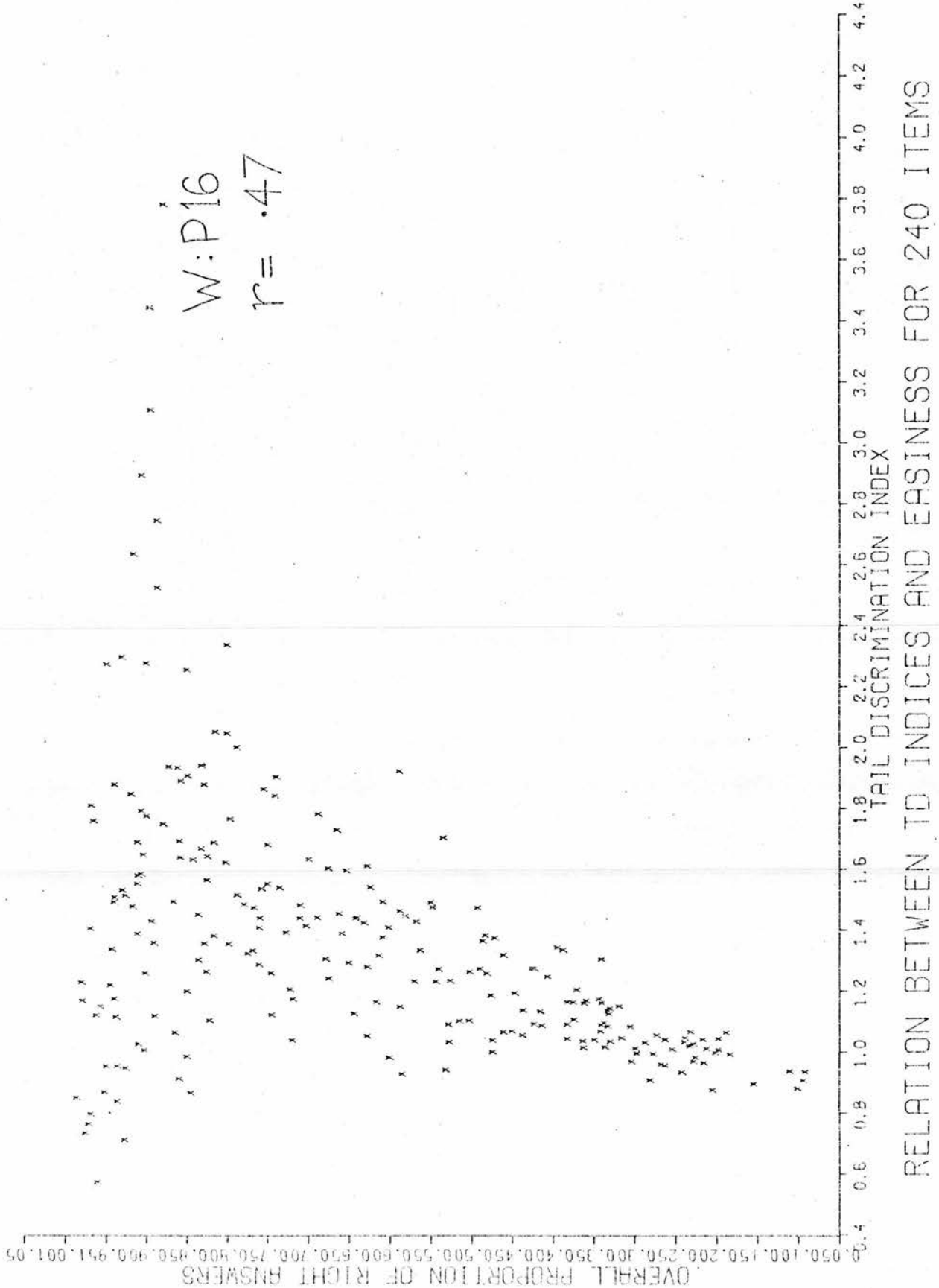

RELATION BETWEEN TD INDICES AND EASINESS FOR 240 ITEMS
W:P16
r= .47
TAIL DISCRIMINATION INDEX
0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4
OVERALL PROPORTION OF RIGHT ANSWERS
.05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60 .65 .70 .75 .80 .85 .90 .95 1.00

FIGURE 26.17

W:P2
r=-.38
RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
TAIL DISCRIMINATION INDEX X10-1
0.24 0.28 0.32 0.36 0.40 0.44 0.48 0.52 0.56 0.60 0.64 0.68 0.72 0.76 0.80 0.84 0.88 0.92 0.96 1.00 1.04
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0

FIGURE 26.18

RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
W:P4
r=-.09
TAIL DISCRIMINATION INDEX
1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0 8.4 8.8 9.2 9.6
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0

FIGURE 26.19

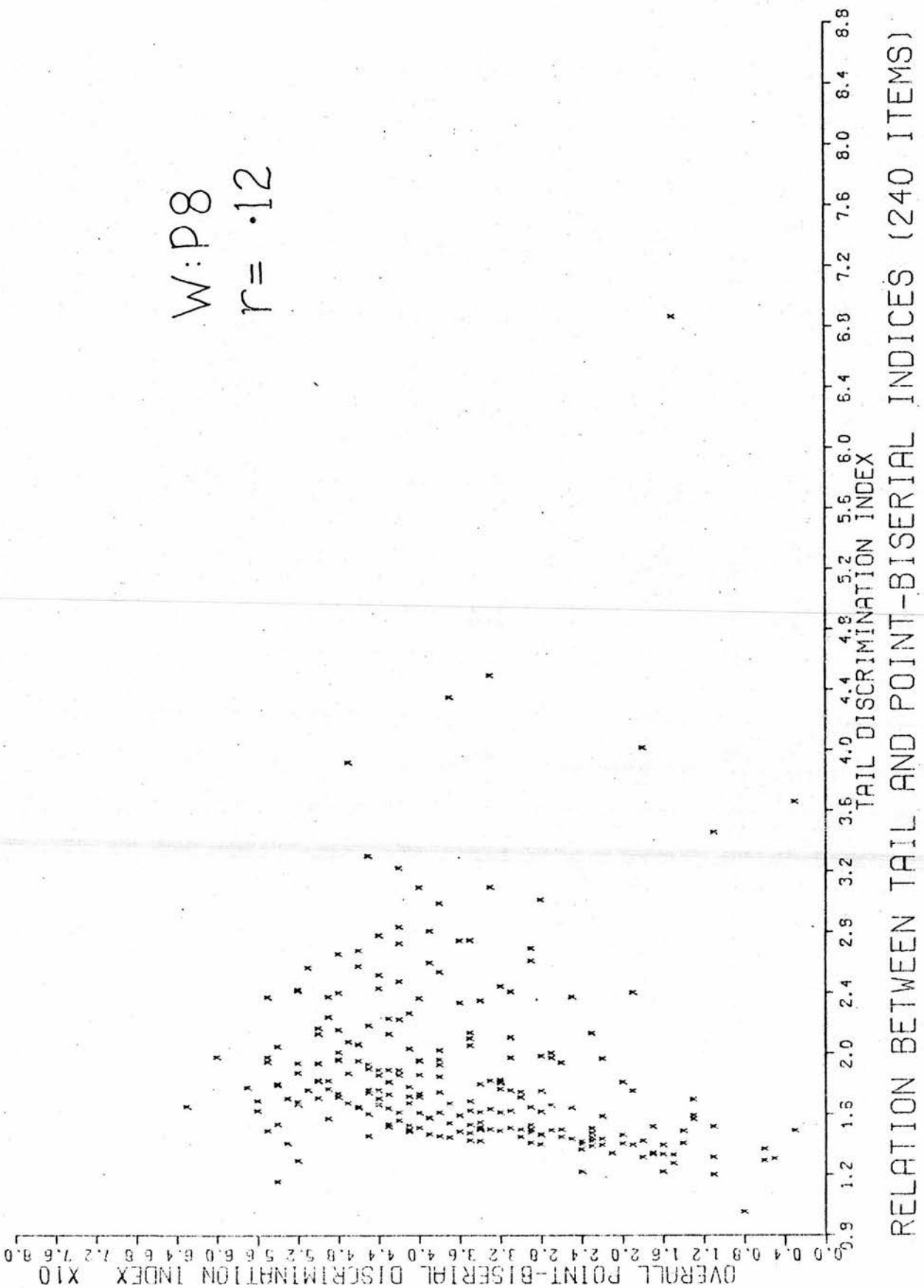

W:P8
r= .12
RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
TAIL DISCRIMINATION INDEX
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10

FIGURE 26.20

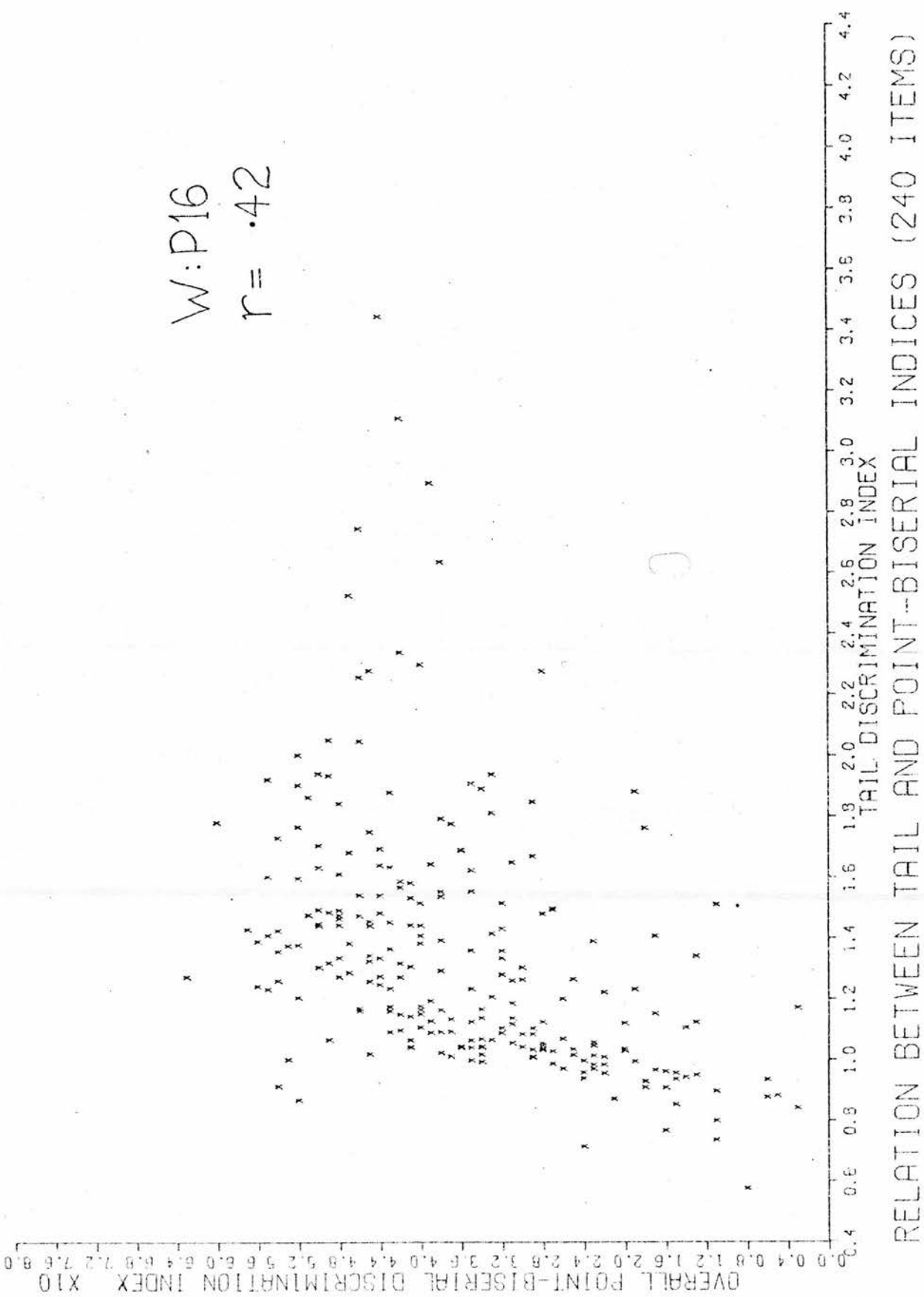

W:P16
r = .42
RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
TAIL DISCRIMINATION INDEX
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10

FIGURE 26.21

R:P2
r = .30
RELATION BETWEEN TD INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX X10-1
0.32 0.36 0.40 0.44 0.48 0.52 0.56 0.60 0.64 0.68 0.72 0.76 0.80 0.84 0.88 0.92 0.96 1.00 1.04 1.08 1.12
TAIL PERCENTILES X10-1
.041 .081 .121 .161 .201 .241 .281 .321 .361 .401 .441 .481 .521 .561 .601 .641 .681 .721 .761 .801 .84

FIGURE 26.22

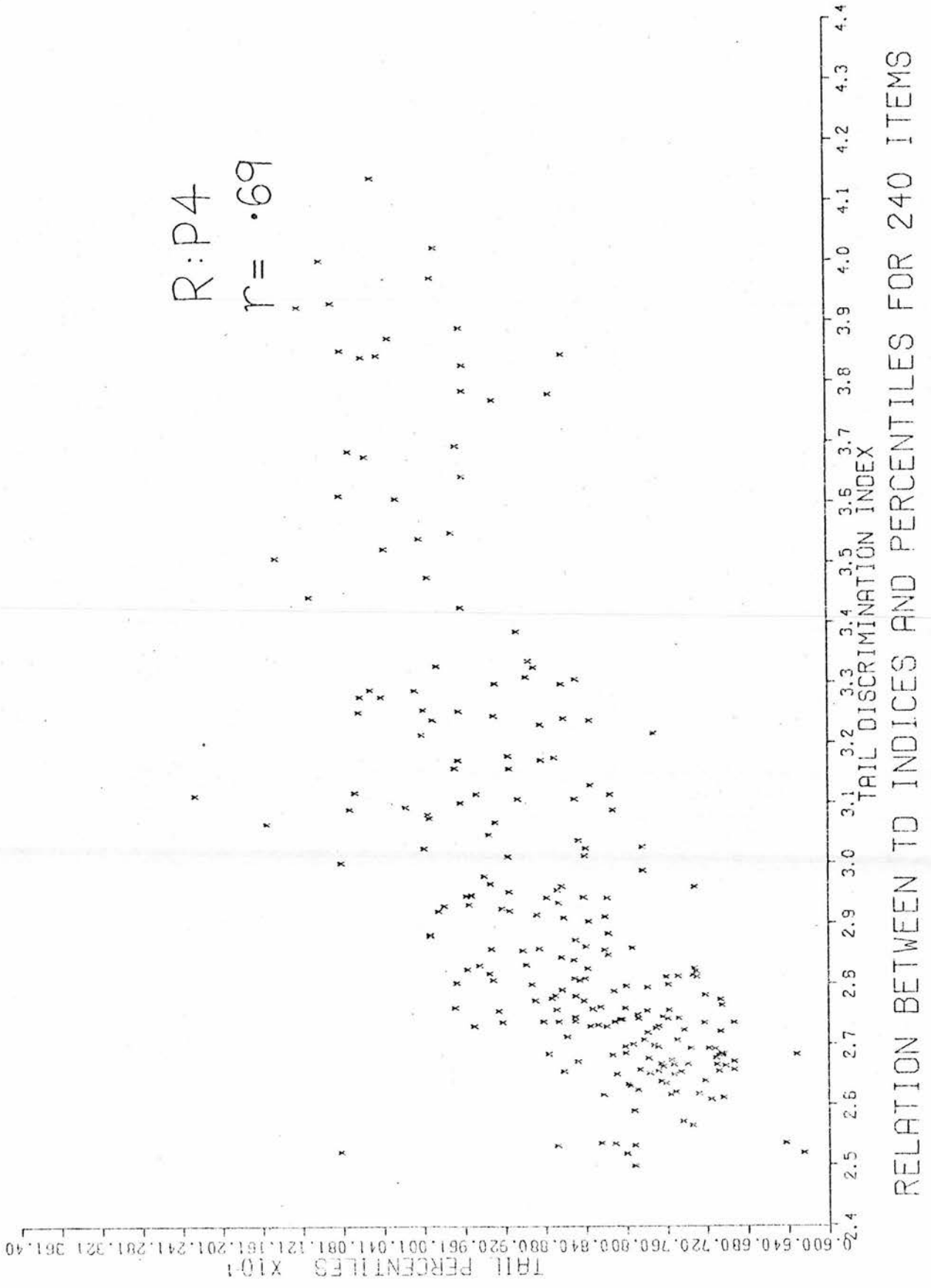

R:P4
r = .69
RELATION BETWEEN TD INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.5 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4
TAIL PERCENTILES X10¹
.600 .640 .680 .720 .760 .800 .840 .880 .920 .961 .001 .041 .081 .121 .161 .201 .241 .281 .321 .361 .40

FIGURE 26.23

R:P8
r= .80
RELATION BETWEEN TD INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6 4.8 5.0 5.2 5.4
TAIL PERCENTILES X10-1
.360 .400 .440 .480 .520 .560 .600 .640 .680 .720 .760 .800 .840 .880 .920 .961 1.001 1.041 1.081 1.121 1.16

FIGURE 26.24

R:P16
r= .88
RELATION BETWEEN TO INDICES AND PERCENTILES FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6
TAIL PERCENTILES X10
0.200 240 280 320 360 400 440 480 520 560 600 640 680 720 760 800 840 880 920 961.00

FIGURE 26.25

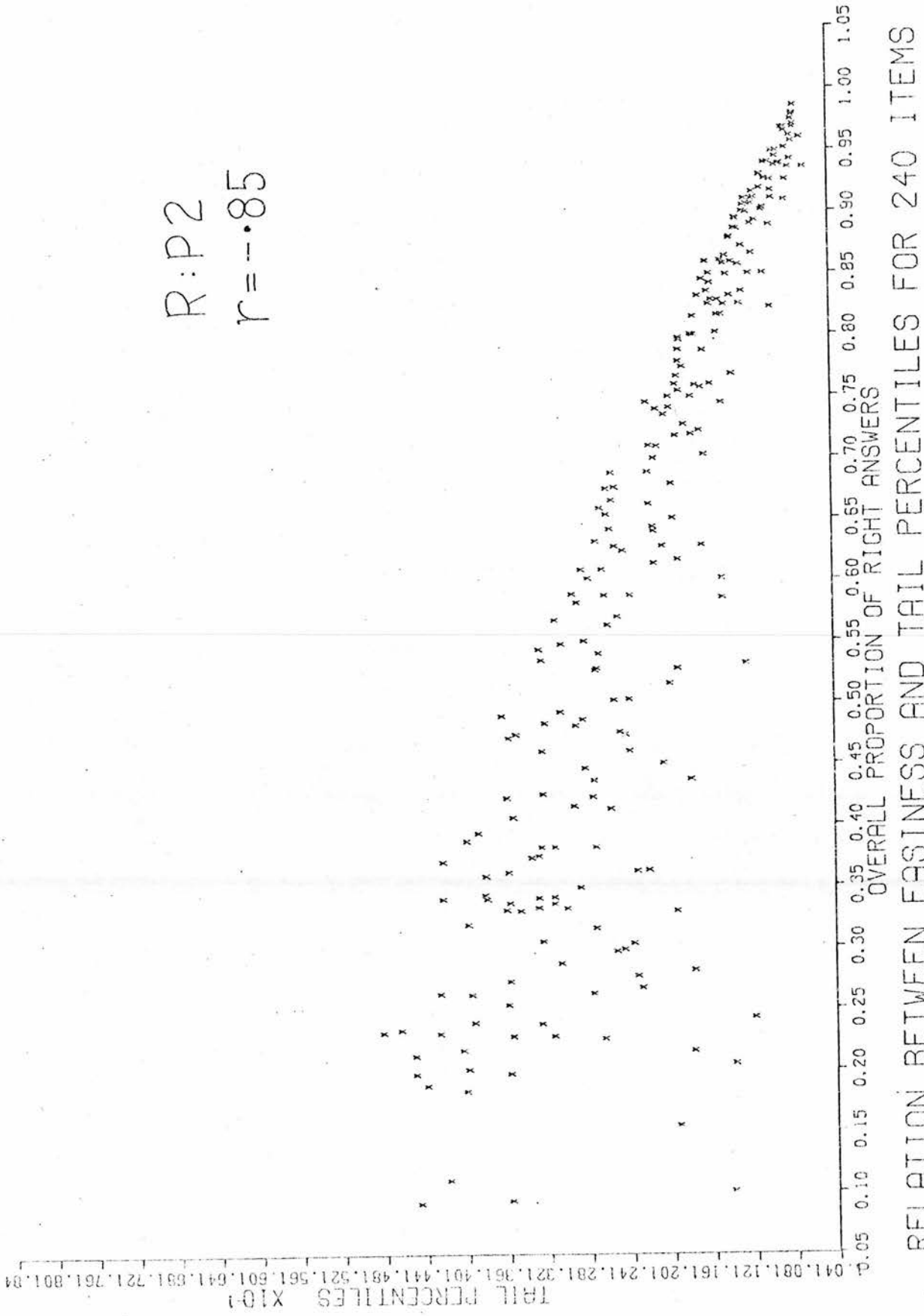

R:P2
r=-.85
RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10-1
041.081.121.161.201.241.281.321.361.401.441.481.521.561.601.641.681.721.761.801.84

FIGURE 26.26

R:P4
r=-.67
RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10-1
.600 .640 .680 .720 .760 .800 .840 .880 .920 .961 .001 .041 .081 .121 .161 .201 .241 .281 .321 .361 .40

FIGURE 26.27

R:P8
r = -.55
RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10
0. 360. 400. 440. 480. 520. 560. 600. 640. 680. 720. 760. 800. 840. 880. 920. 961. 001. 041. 081. 121. 16

FIGURE 26.28

R:P16
r=-.39
RELATION BETWEEN EASINESS AND TAIL PERCENTILES FOR 240 ITEMS
OVERALL PROPORTION OF RIGHT ANSWERS
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05
TAIL PERCENTILES X10-1
0. 200. 240. 280. 320. 360. 400. 440. 480. 520. 560. 600. 640. 680. 720. 760. 800. 840. 880. 920. 961.50

FIGURE 26.29

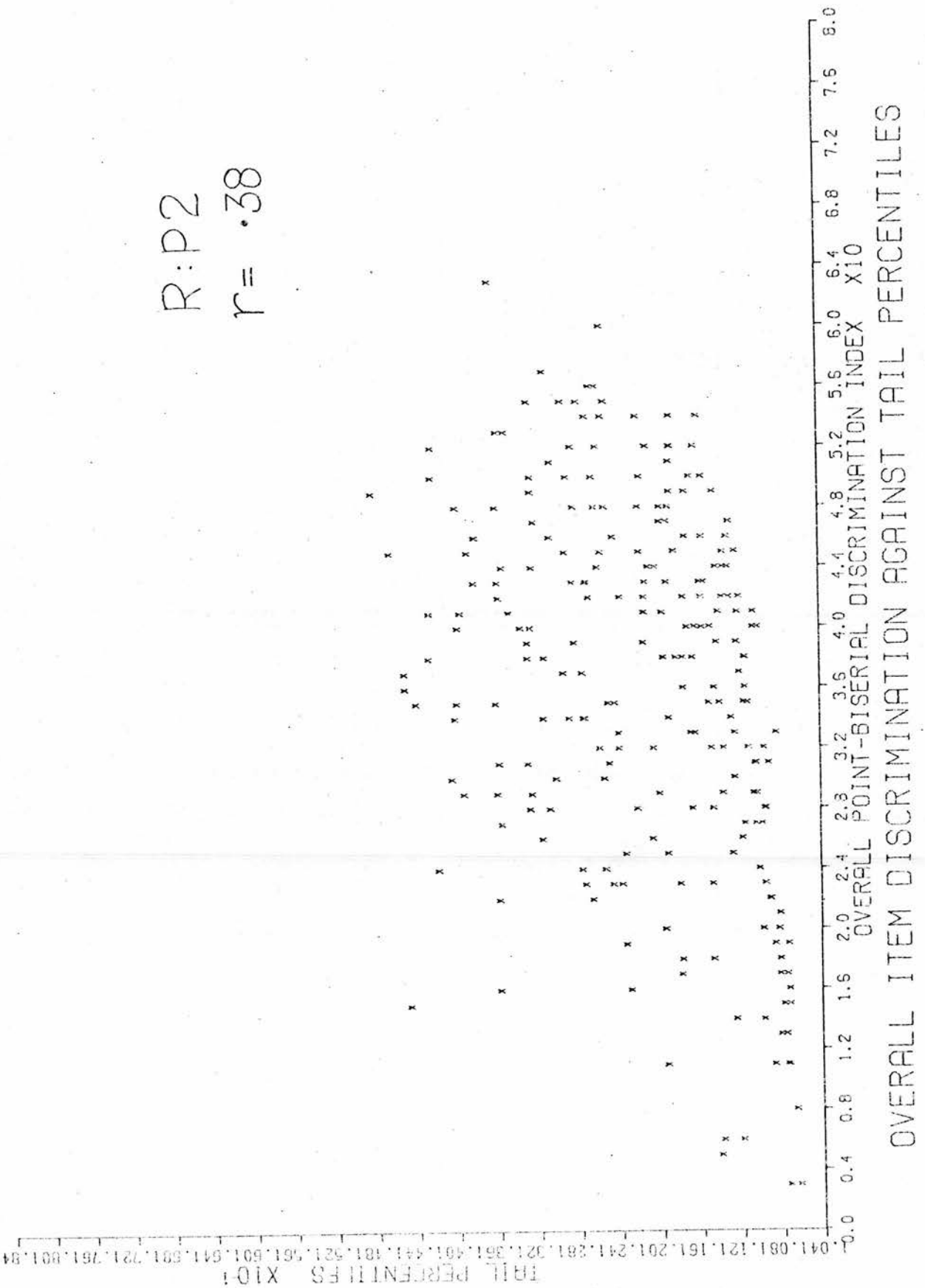

R:P2
r= .38
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
TAIL PERCENTILES X10

FIGURE 26.30

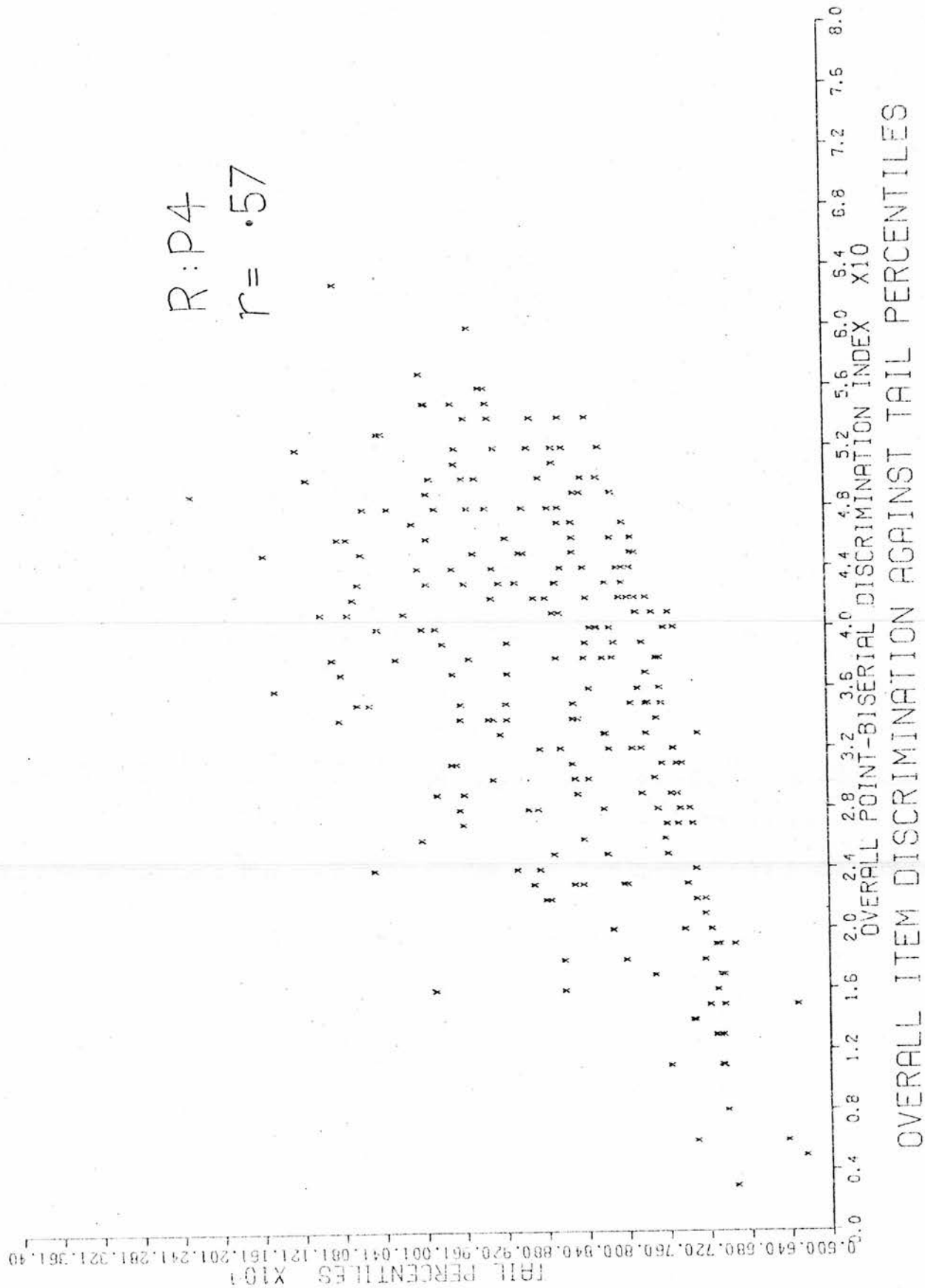

R:P4
r= .57
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
TAIL PERCENTILES X10
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0
6.4
6.8
7.2
7.6
8.0
0.
600.
640.
680.
720.
760.
800.
840.
880.
920.
961.
001.
041.
081.
121.
161.
201.
241.
281.
321.
361.
40

FIGURE 26.31

R:P8
r= .67
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0
TAIL PERCENTILES X10
0.0 .360 .400 .440 .480 .520 .560 .600 .640 .680 .720 .760 .800 .840 .880 .920 .961 1.001 1.041 1.081 1.121 1.16

FIGURE 26.32

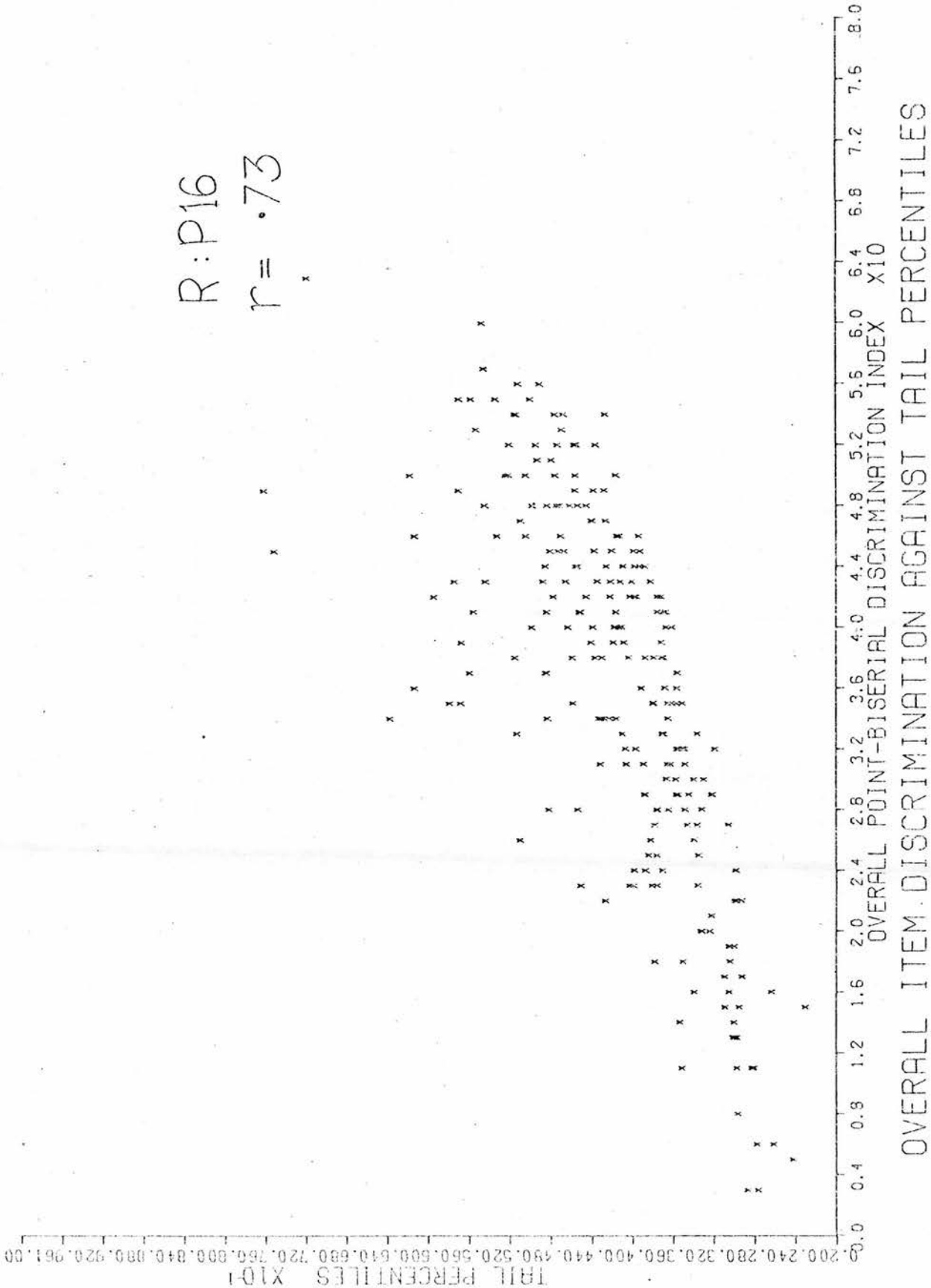

R:P16
r = .73
OVERALL ITEM DISCRIMINATION AGAINST TAIL PERCENTILES
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
TAIL PERCENTILES X10
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0
6.4
6.8
7.2
7.5
8.0
0.00
0.40
0.80
1.20
1.60
2.00
2.40
2.80
3.20
3.60
4.00
4.40
4.80
5.20
5.60
6.00
6.40
6.80
7.20
7.60
8.00
8.40
8.80
9.20
9.60
10.00

FIGURE 26.33

R:P2
r=-.52
RELATION BETWEEN TO INDICES AND EASINESS FOR 240 ITEMS
TAIL DISCRIMINATION INDEX X10⁻¹
0.32 0.36 0.40 0.44 0.48 0.52 0.56 0.60 0.64 0.68 0.72 0.75 0.80 0.84 0.88 0.92 0.96 1.00 1.04 1.08 1.12
OVERALL PROPORTION OF RIGHT ANSWERS
.050 .100 .150 .200 .250 .300 .350 .400 .450 .500 .550 .600 .650 .700 .750 .800 .850 .900 .951 .001 .05

FIGURE 26.34

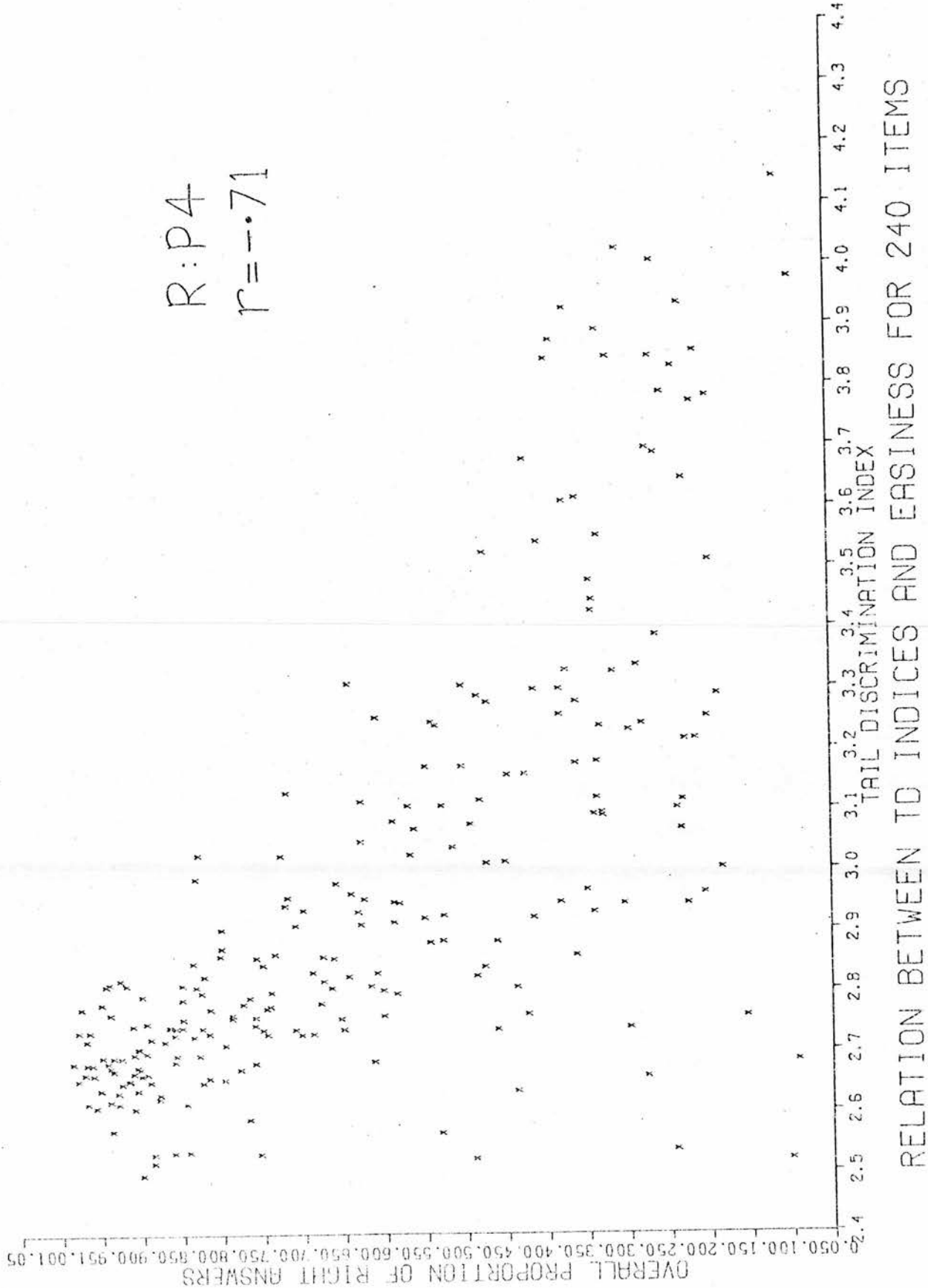

R:P4
r=-.71
RELATION BETWEEN TD INDICES AND EASINESS FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
OVERALL PROPORTION OF RIGHT ANSWERS

FIGURE 26.35

RELATION BETWEEN TD INDICES AND EASINESS FOR 240 ITEMS
R:P8
r=-.66
OVERALL PROPORTION OF RIGHT ANSWERS
.050 .100 .150 .200 .250 .300 .350 .400 .450 .500 .550 .600 .650 .700 .750 .800 .850 .900 .951 1.05
TAIL DISCRIMINATION INDEX
1.4 1.5 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6 4.8 5.0 5.2 5.4

FIGURE 26.36

R:P16
r=-.50
RELATION BETWEEN TD INDICES AND EASINESS FOR 240 ITEMS
TAIL DISCRIMINATION INDEX
OVERALL PROPORTION OF RIGHT ANSWERS
.05 .001 .95 .90 .85 .80 .75 .70 .65 .60 .55 .50 .45 .40 .35 .30 .25 .20 .15 .10 .05 0.0
0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6

FIGURE 26.37

R:P2
r=-.48
RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
TAIL DISCRIMINATION INDEX X10⁻¹
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10

FIGURE 26.38

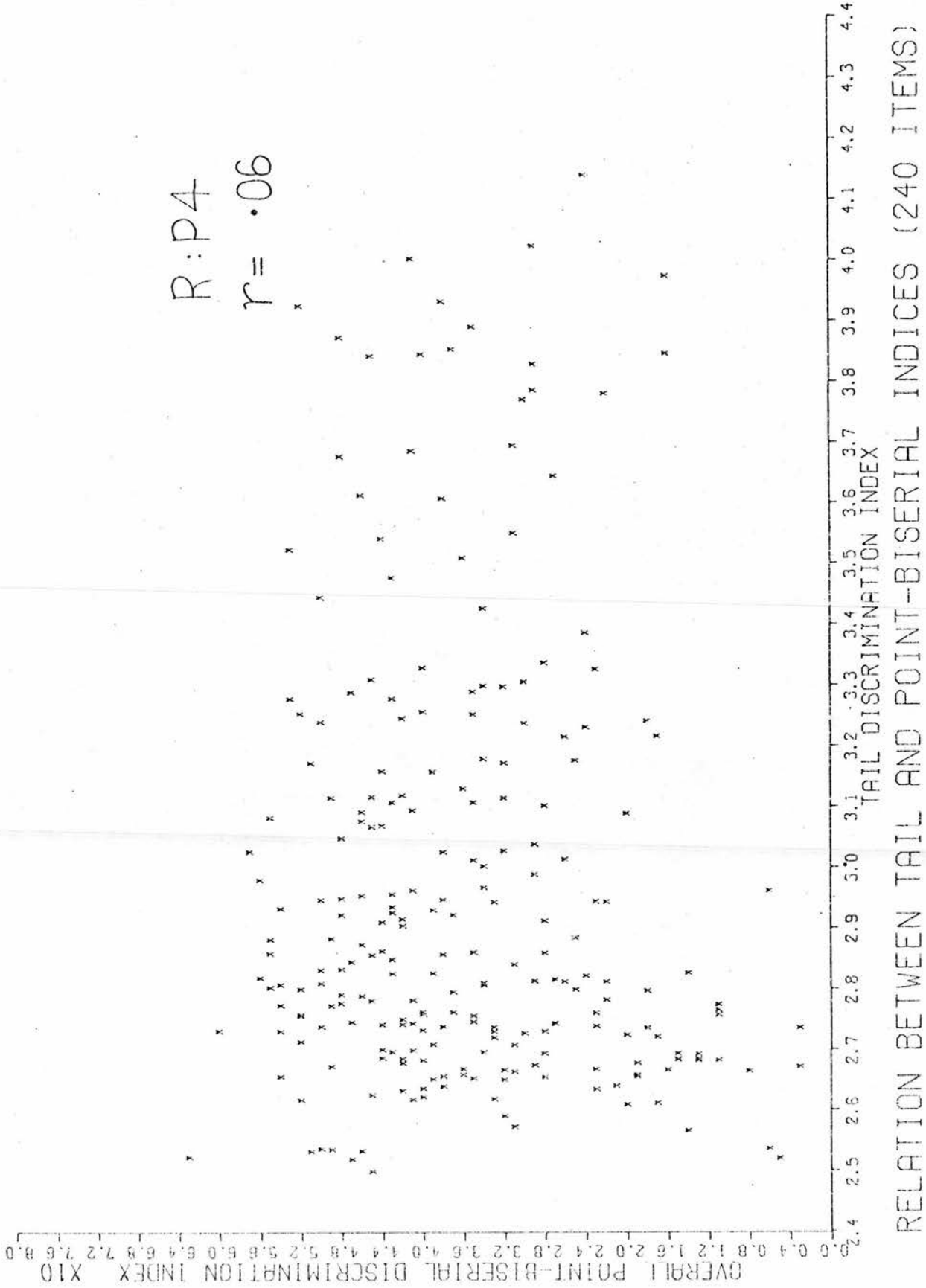

R:P4
r= .06
RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
TAIL DISCRIMINATION INDEX
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10

FIGURE 26.39

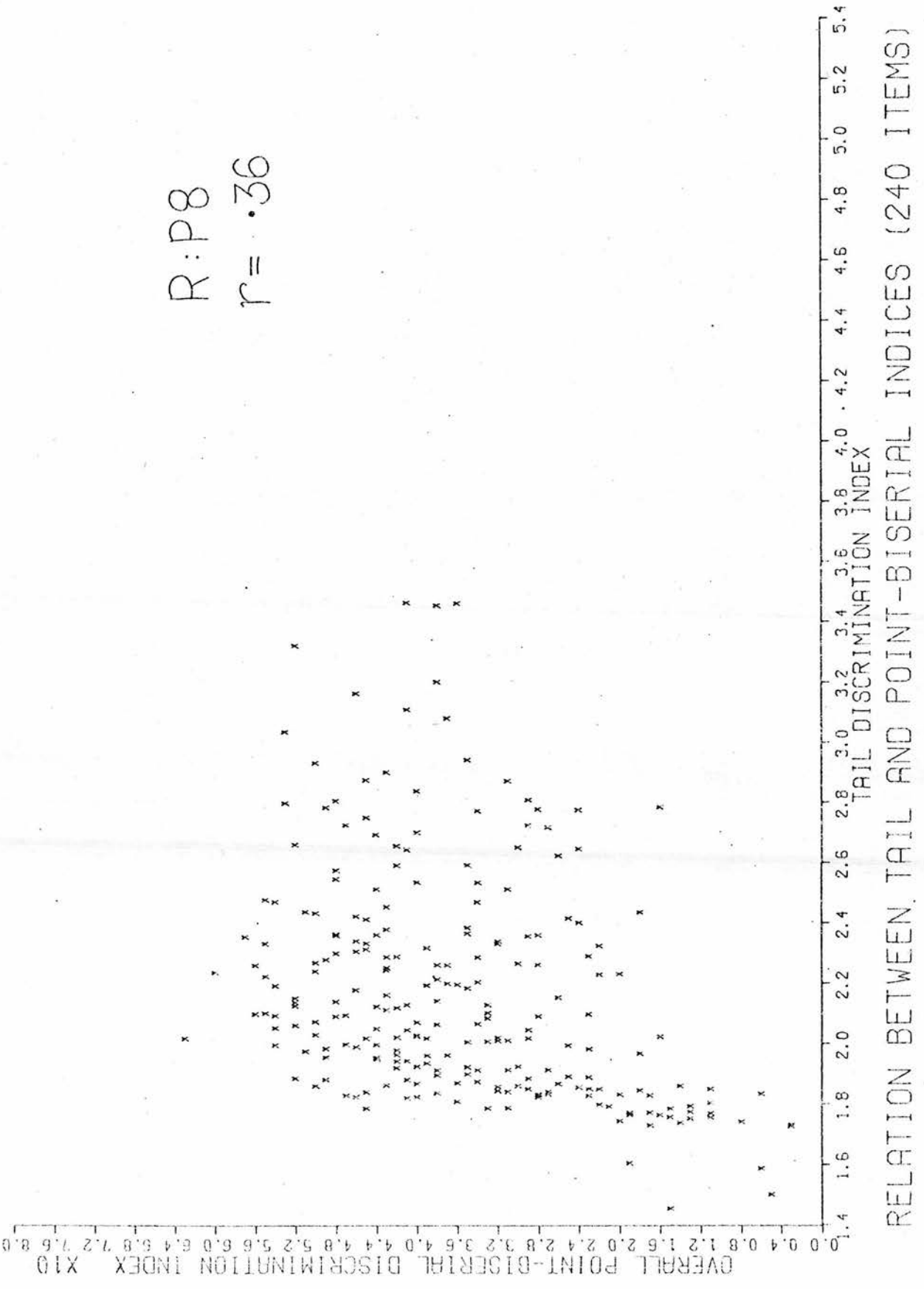
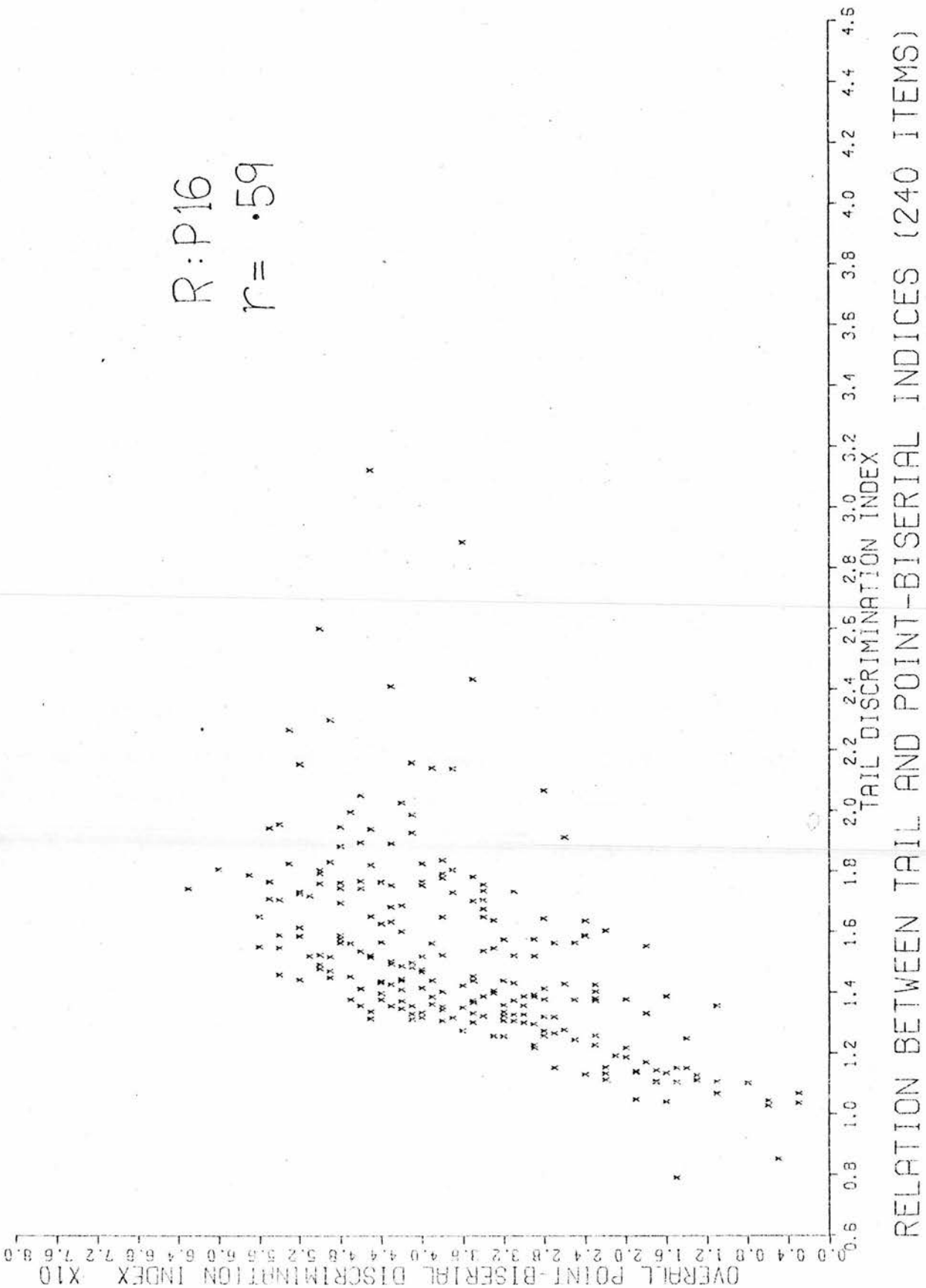

R:P8
r = .36
RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
TAIL DISCRIMINATION INDEX
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10

FIGURE 26.40


RELATION BETWEEN TAIL AND POINT-BISERIAL INDICES (240 ITEMS)
R:P16
r= .59
TAIL DISCRIMINATION INDEX
0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6
OVERALL POINT-BISERIAL DISCRIMINATION INDEX X10
0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.5 6.0 6.4 6.8 7.2 7.6 8.0

and few exhibit the linearity or homoscedasticity that would make a Pearson-r fully appropriate. Pearson-r has, however, been computed for the plots to be used simply as a comparative, descriptive statistic. The possibility of chance success is responsible for the general quantitative asymmetry between right- and wrong-curves, but the pattern of relationships is qualitatively similar.

The point-biserial correlation coefficient is known to be markedly affected by extreme splits; for very easy or very hard items its value is low. For such items PMD may also be low because of the population distribution's constraining influence. This link is behind some features of the plots. The more extreme right/wrong splits will also result in tail indices being based on smaller or even much smaller samples with consequently greater sampling error.

Commenting now on the individual pairings and using the right-curve (R) and wrong-curve (W) labelling of the plots.

Plot 1 P-values against PMD-values.

The difference in the direction of the R and W relationships results from their concern with opposite tails. Quantitatively the P2 plots stand apart from the others and display looser relationships. While it appears possible for low (good) PMD-values to occur for all W Tail Location bands, it may not be possible for central R Tail Locations to obtain similarly good Tail Discrimination.

Plot 2 P-values against overall easiness.

Irrespective of set lower Tail Locations are associated with easier questions. For P2, the median, the relationship is expectedly substantial.

Plot 3 P-values against overall discrimination.

The closest relationship is for P16. The more central this tail-end location the higher the overall discrimination.

Plot 4 PMD-values against overall easiness.

Low PMD-values occur over the full range of easiness.

Plot 5 PMD-values against overall discrimination.

Clearly PMD2 and PMD16 have a degree of counter-relationship as the scatterplots change direction from one extreme to the other. Plausibly a blunt PMD2 implies a tapered PMD16. The relationship is not close but for PMD8 and PMD16 better Tail Discrimination is associated with poorer overall discrimination.

The relationship between the tail indices and conventional item characteristics is complex: they inter-relate through a number of influences. Subject to the limitations on interpretation already mentioned Tables 6 and 7 present the Pearson-r values between tail indices and conventional characteristics and also between the several tail indices.

In selecting the definitive P-value and PMD-value for the present research the P2 and PMD2 values based on the median are not strictly contenders; P2 is hardly a tail value, and while PMD2 is a half-distribution value it is rather too often the odd-one-out to be acceptable as representative. At the other extreme P16 and PMD16 are dependent on the most improbable parts of the distribution and will have high sampling errors: unless these values show especial virtues more stable indices are to be preferred.

All the P-values are highly inter-related: if P2 is eliminated as too similar to overall item easiness, then P8 is the most represent-

TABLE 6 Pearson product-moment correlations between tail indices and conventional item characteristics.

	Conventional Item Statistics	
	Easiness	Point-biserial
Tail Location		
Wrong-curve		
P2	-·84	-·34
P4	-·75	-·42
P8	-·67	-·49
P16	-·59	-·56
Right-curve		
P2	-·85	·38
P4	-·67	·57
P8	-·55	·67
P16	-·39	·73
Tail Discrimination		
Wrong-curve		
PMD2	-·52	-·48
PMD4	-·71	·06
PMD8	-·66	·36
PMD16	-·50	·59
Right-curve		
PMD2	·30	-·38
PMD4	·54	-·09
PMD8	·51	·12
PMD16	·47	·42

TABLE 7 Pearson-r correlations between tail indices of Tail Location (P-values) and Tail Discrimination (PMD-values). (Decimal points omitted)

Wrong-curve

	P2	P4	P8	P16	PMD2	PMD4	PMD8	PMD16
P2		94	86	77	-23	-61	-63	-63
P4			96	87	09	-54	-68	-70
P8				94	27	-31	-62	-72
P16					37	-07	-33	-67
PMD2						53	10	-16
PMD4							72	32
PMD8								61
PMD16								

Right-curve

	P2	P4	P8	P16	PMD2	PMD4	PMD8	PMD16
P2		91	82	67	30	77	83	75
P4			96	83	-10	69	89	88
P8				94	-29	47	80	93
P16					-44	19	57	88
PMD2						43	05	-27
PMD4							83	44
PMD8								81
PMD16								

ative of the remaining three - being highly related to both P4 and P16 because of its middle position. P8 is considered sufficiently tailish to index a difficulty level localised to the biting edge of an item in helping the derived distribution to converge.

Among the PMD-values PMD2 and PMD16 are negatively related. PMD8 and PMD16 both appear susceptible to the boundary effect of the population distribution in their relationship with P-values (see, for example Figures 26.4, 26.5, 26.8 and 26.9). For the right-curve PMD8 and PMD16 are too highly related to P8. PMD4 is the preferred choice, avoiding the worst of the above faults and having reference to a sizable tail so that an item with a low PMD4 may take a bite rather than a nibble out of the derived distribution. PMD2 would perhaps be in danger of being too central an index and thus allow the derived distribution to become too tapered - this danger being underlined by its negative relationship to PMD16.

The two indices selected, P8 and PMD4, are related only moderately (about ±0·4). P8 relates about ±0·6 with overall easiness and with overall discrimination. PMD4 relates similarly to easiness but only about ±0·1 with discrimination. There is thus room for the two indices to be independently useful: whether they are so or not will emerge after they have been used to select the item library and to control question choices from the library within the tailored testing procedure.

B. Selecting the item library and a further comparison of tail and conventional item indices

The item library was chosen in stages. (The basic data are the P-values

and PMD-values in Annex X.) First the item pool was screened to establish which of the 240 items were eligible for the W-set or R-set on a Tail Discrimination standard alone. For the W-set this standard was a value of PMD4 less than 2·45, and for the R-set it was PMD4 less than 2·90. It can be seen roughly from Table 4 (p. 138) that these values will eliminate about a half of the item pool in each case.

After screening on Tail Discrimination 134 items remained eligible for the W-set and 137 items for the R-set. Most items were eligible for one set or the other. 54 items were eligible for both sets and only 23 for neither set. Of these 23 all but two were acceptable[1] for conventional test purposes.

Table 8 gives the distributions by Tail Location, P8, of the eligible items. Compared with the full item pool (given in Table 3) the distributions remain similar.

The final selection of the item library from the eligible items had the following aims,

i. to choose four items for the W-set and four for the R-set (with possible overlap) from each of the twelve 20-item tests. (See pp. 139-140.)

ii. to achieve as even a spread as possible of Tail Location, P8.

iii. to choose items with the lowest Tail Discrimination, PMD4.

Aim i. was mandatory, and aim iii. had lowest priority as the eligible items were already screened on PMD4.

The distribution of the chosen library items by Tail Location, P8,

---

1. Against the criteria of overall easiness being in the range 40% to 90%, and overall discrimination being not less than 0·3.

TABLE 8. The frequency distribution by verbal attainment band of Tail Location (indexed by P8) for items average or better on a Tail Discrimination eligibility criterion, and for the library items.

Attainment Band[a]	Eligible Items		Library Items	
	W[b]	R[b]	W	R
1				
2				
3				
4		5		1
5		58		21
6		53		13
7		18		10
8	1	2	1	2
9	1	1	1	1
10	7		6	
11	10		6	
12	14		8	
13	16		7	
14	25		6	
15	49		7	
16	11		6	
17				
18				
19				
	134	137	48	48

Notes:

a The nominal bands take in values down to 0·5 below and up to (but not including) 0·5 above the given band.

b The W and R columns refer to the wrong- and right-curves (as indicated in Figure 16).

is also given in Table 8. Ten items are common to both the W-set and the R-set, hence the full item library used here in assembling the response banks consists of 86 items in their two (overlapping) sets of 48.

The derived distributions for these 86 items are illustrated in Figures 27.1 to 27.8 and in Annex XIII. The Figures also identify the items and indicate to which set they belong as explained at the start of Figure 27. (The data for these distributions are those of Annex IX.)

Finally Figure 28 plots all the 240 items of the pool on a scatterplot of overall easiness (indexed by percentage of recruits answering correctly) against overall discrimination (indexed by point-biserial correlation). On this plot the library items are distinguished by symbols indicating to which set they belong. The open box superimposed on the plot indicates the items acceptable for a conventional test using the criteria previously defined.

For the conventional test 123 items are acceptable (51%), but these include only 49 of the 86 library items (57%). Hence conventional acceptability is only marginally related here to the criteria for library selection. 37 of the library items lie outside the box and would not be conventionally acceptable.

In Figure 28 the library items tend to occupy positions round the outer periphery of the scatterplot, however, this is a tendency only as the relationship with overall discrimination for an easiness level is imperfect. The conventional acceptable easiness range centres on 65% because the items are 5-option multiple-choice. Notwithstanding this asymmetry the library selection procedures have successfully ensured an approximately equal number of library items outside the box at its

FIGURE 27

Graph plots of the item/population derived distributions for eight library items. (Plots for the remaining library items appear in Annex XIII.)

At the foot of each graph the individual item is identified in the usual way and the letter W or R appended indicates membership of the W-set or R-set respectively.

Figures 27.1 to 27.8 follow:-

FIGURE 27.1

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
1/1 R
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10⁻¹
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
NORMATIVE PROPORTION RIGHT (DASHED LINE)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 27.2

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
1/5 R
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10⁻¹
0.0 -0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
NORMATIVE PROPORTION RIGHT (DASHED LINE)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 27.3

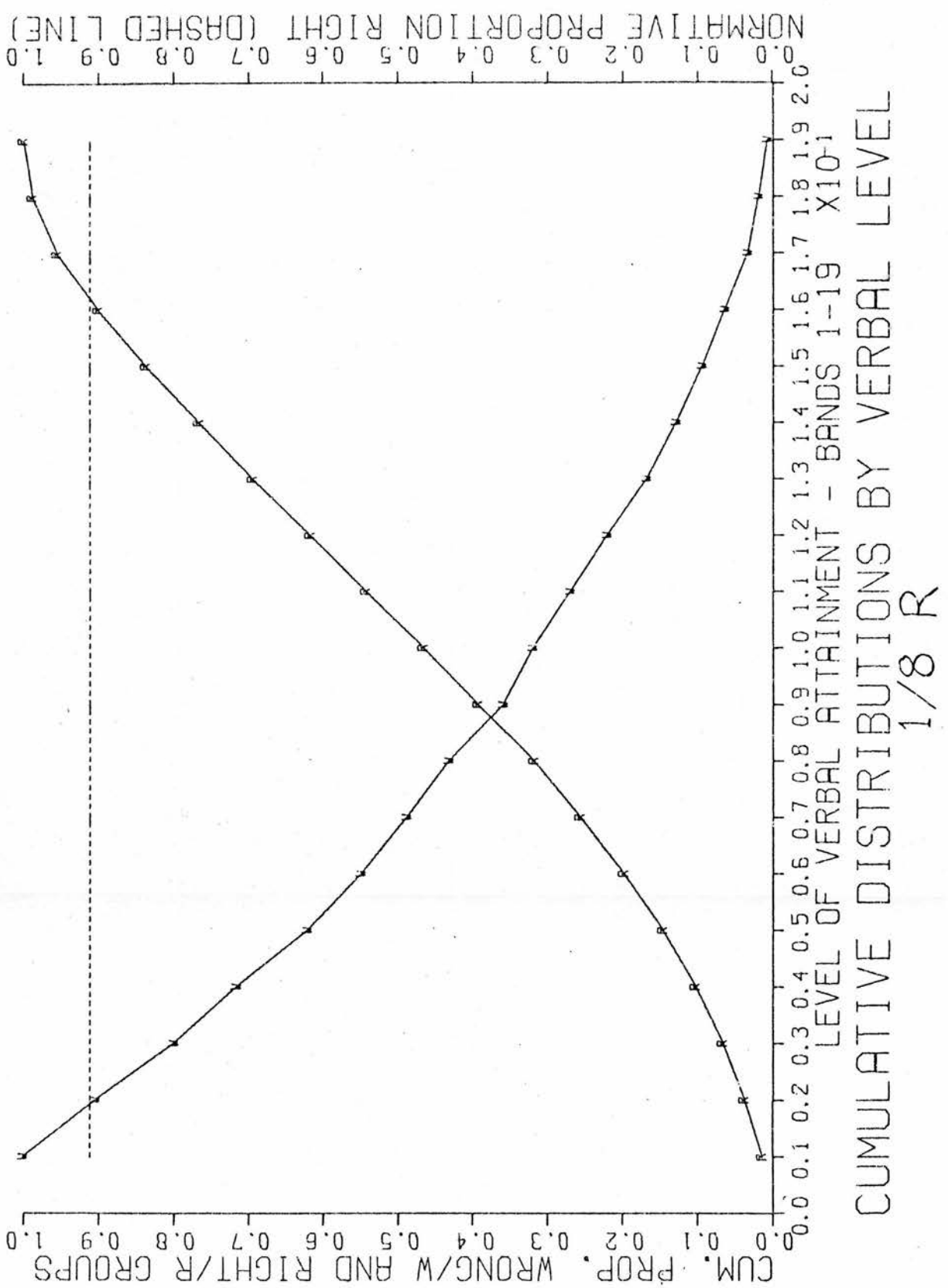

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
1/8 R
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
NORMATIVE PROPORTION RIGHT (DASHED LINE)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 27.4

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
1/10 W
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10⁻¹
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
NORMATIVE PROPORTION RIGHT (DASHED LINE)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 27.5

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
1/12 W
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
NORMATIVE PROPORTION RIGHT (DASHED LINE)

FIGURE 27.6

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
1/13 WR
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
NORMATIVE PROPORTION RIGHT (DASHED LINE)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

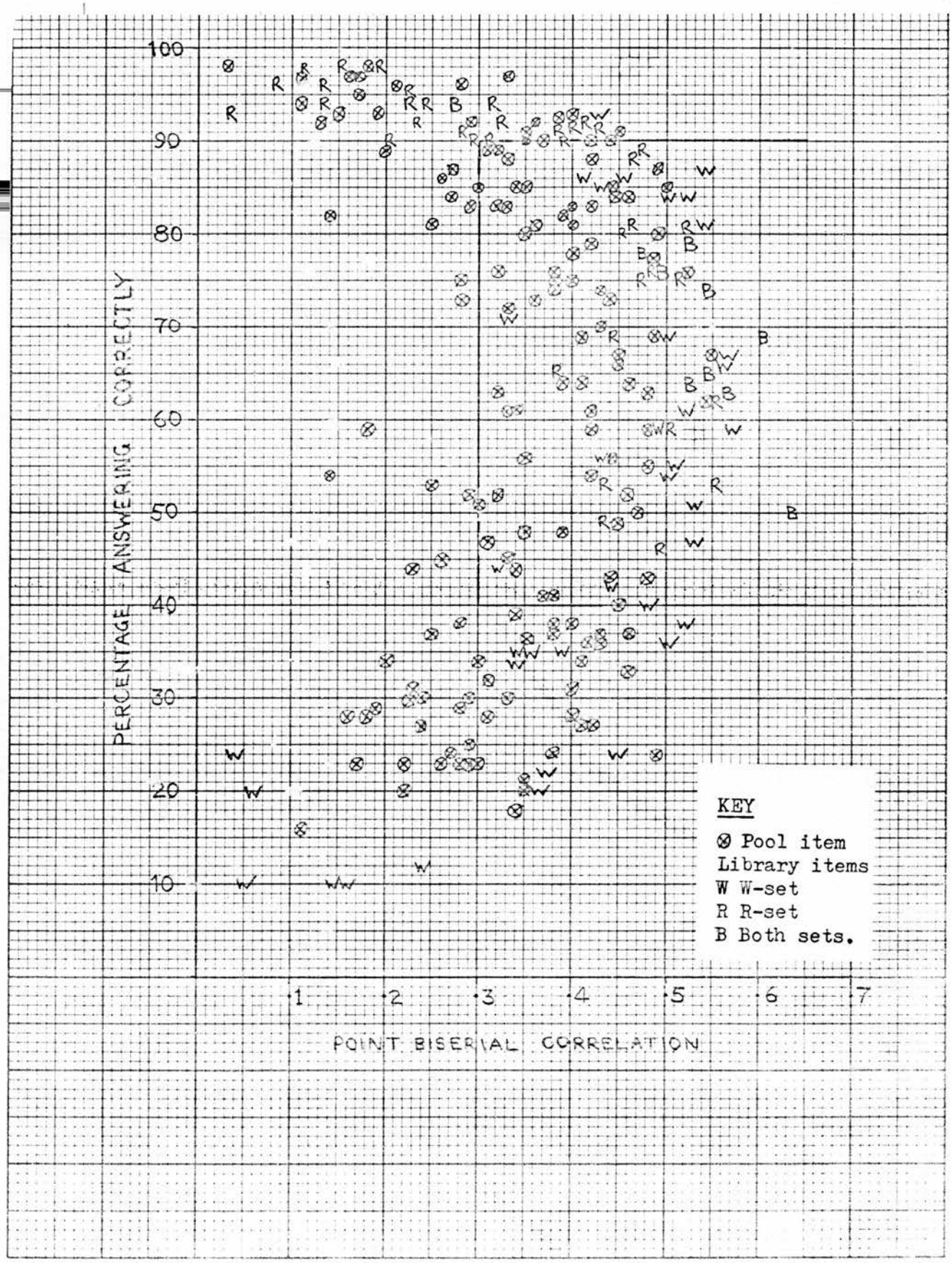
FIGURE 27.7

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
1/18 W
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
NORMATIVE PROPORTION RIGHT (DASHED LINE)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 27.8

CUMULATIVE DISTRIBUTIONS BY VERBAL LEVEL
2/1 R
CUM. PROP. WRONG/W AND RIGHT/R GROUPS
NORMATIVE PROPORTION RIGHT (DASHED LINE)
LEVEL OF VERBAL ATTAINMENT - BANDS 1-19 X10-1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 28 A scatterplot of conventional item indices for the verbal pool showing the library items.



easier and harder limits. It is to be expected that items for tailored testing should cover the full easiness range: the conventional test must be constructed to suit the middle range or modal recruit, whereas it is especially for the non-modal recruit that tailored tests will show most benefit.

The most difficult items are in the W-set, and the easiest are predominantly in the R-set. This is partly the nature of the tailoring process - the R-set has to be able to establish lower bounds for recruits of low attainment, and conversely for the W-set - but it also represents a deficiency which could in part be made good by items not sufficiently available in the item pool. It is this structural imbalance which the tailoring process deals with by means of what has been described as a tail change.

Chapter 8: Results III presents and discusses the findings from the use of the item library and its response banks in the simulation of the proposed tailored testing procedure.

7. RESULTS II AND DISCUSSION: A TEST OF THE ASSUMPTION OF LOCAL INDEPENDENCE

Following the method given in Chapter 5.C, PROGRAM 4 (introduced on p.142 and given at Annex XI) computed probabilities and Chi-Square for 144 item pairs for 6 wide attainment levels and 6 narrow attainment levels. The 144 pairs represented all possible pairings for the item library within sets within tests.

The assumption of local independence would not necessarily require the performance on two items to be independent for the wide attainment levels. Wide here means a spread of three verbal attainment bands and this may be insufficiently local. Local strictly refers to one point only on the continuum of attainment. How narrowly local needs to be interpreted in a particular application depends on the penalties resulting from the progressive untenability of local independence as it is more broadly assumed.

However, if it can be shown that local independence is a realistic assumption for a wide attainment level then it would follow a fortiori for the narrow levels. Hence in testing the assumption here attention is principally directed at the wide attainment levels. This emphasis is appropriate not only because the wide level is the more demanding test of the assumption. Response banks have been assembled for wide as well as narrow attainment levels in order to increase the number of independent simulated tests possible, so that it is relevant to know what departure from local independence (if any) such banks may display. It is also the case that the larger samples afforded by the wide attainment definition make for readier statistical checks of the assumption.

The main analysis is carried out on samples which give rise to item-pair 2-by-2 contingency tables based on 40 or more joint item responses. The source of each joint response is a different recruit and hence all the table entries are independent. The minimum sample size of 40 is judged sufficient to support a Chi-Square approximation of the multinomial distribution except for extreme splits. There is no generally agreed minimum sample size because of the additional complication of the relative proportions associated with the two splits being cross-tabulated. (A minimum sample size of 40 is suggested by Siegel[1].) When any expected cell frequency drops to a value below 5 Chi-Square evaluations may well become unduly inflated.

Annex XIV gives the detailed tabulations of joint probabilities and Chi-Square for the library item pairings at the selected attainment levels. These were wide and narrow band levels 3, 6, 9, 12, 15 and 18. A full explanation of the tabulations is given at the start of the Annex.

In presenting the results the W-set and R-set of library items are dealt with separately because of one point of difference which arises.

Local independence in the W-set items

Of the 432 contingency tables describing joint item performance for the wide attainment levels (72 item pairs x 6 levels), 284 had a sample size of 40 or greater (and with no expected cell frequencies of zero). On first inspection 27 of the 284 Chi-Square values were

---

1. Siegel, S. Nonparametric Statistics for the Behavioural Sciences: New York, McGraw-Hill, 1956.

larger than the 5% probability value. That is, at the 5% wrong-rejection risk level 27 values were significantly larger than would be expected on the null hypothesis of independence. On further inspection five of these 27 values had the following cell frequencies indicating extreme splits:-

2-by-2 table

b	d
a	c

Cell frequencies

a	b	c	d
0	1	1	76
0	2	1	68
0	1	2	68
54	1	3	0
0	1	1	39

The interpretation of the Chi-Square values for these items is highly suspect because of the low expected cell frequencies involved. A Fisher exact probability test was carried out on these five tables and gave non-significant results in each case (again at the 5% level).

Hence we are left with 22 significant Chi-Square values out of 284. This is 7·7% of significant values at the 5% risk level. The small excess of rejections of the null hypothesis over the expected level indicates that even if the small apparent excess is genuine and repeatable the assumption of local independence holds very substantially. For comparison, at the opposite extreme 16 Chi-Square values (5·6%) were so low as to indicate an improbably good fit to the independence assumption at the 95% level.

The same item pairs are compared at several attainment levels. If, where independence does not hold, this is related to the content

of the items, then rejection of the null hypothesis might be expected to occur repeatedly for the same item pair at different attainment levels. This generally did not occur - and this lends further support to the acceptance of the assumption of local independence - but two item pairs (both in the same test and involving three items) did show dependence for more than two attainment levels. The association was positive in each case. Between them these two item pairs account for 7 of the 22 significant instances. Inspection of the three items showed they had no content obviously in common. The best hypothesis of the writer (but not strongly proposed) is that the dependence has to do with the types and success of the distractors, or wrong options, offered in these items. It is perhaps not without significance that these three items occur in the W-set. No similar instances occur in the R-set and this is the point of difference referred to earlier.

Local independence in the R-set items

More of the R-set items gave extreme splits with expected cell frequencies of zero. This could be anticipated from Figure 28 (p.199) where pairings between the easiest items would be expected to give this result. Hence there was a smaller number of eligible contingency tables. 253 of the 432 tables were eligible and had sample sizes of 40 or more. First inspection showed 35 Chi-Square values to be significantly large at the 5% level. Further inspection showed that 18 of these 35 contingency tables had the following cell frequencies indicating extreme splits:-

2-by-2 table			Cell frequencies			
b	d		a	b	c	d
a	c		0	3	1	78

Cell frequencies			
a	b	c	d
0	1	3	78
0	2	1	75
0	3	1	74
0	4	1	75
0	1	1	65
0	1	1	58
0	1	1	58
0	1	1	58
0	1	2	82
0	1	1	83
1	1	0	83
0	1	1	74
0	1	2	73
0	1	2	60
0	2	1	38
0	2	1	38
0	1	1	39

Fisher exact probability tests on these tables again indicated non-significance in all cases.

Hence there are 17 significant values out of 253 cases. This is 6·7% of values at the 5% risk level. For comparison 11 values (4·4%) show an improbably good fit to the null hypothesis at the 95% level. Again local independence is seen to hold very substantially (if not completely).

Apart from attainment band 3 (which had too few recruits) all of

the attainment levels contributed in some measure to the qualifying contingency tables examined above. As a safeguard the Fisher exact probability test was used to examine 40 contingency tables for attainment band 3 - 20 from each of the sets. None of these tests gave significant results at the 5% level. For the small samples involved very gross deviations from the null hypothesis are required for its rejection.

The results reported give extremely substantial support to the assumption of local independence - even when local means 3 bands wide. In particular they endorse the viability of the wide response banks that are to be used for the simulations in the next chapter.

8. RESULTS III AND DISCUSSION: REAL-DATA SIMULATION OF THE PROPOSED TAILORED TESTING METHOD.

A. Introduction

In this Chapter the central issue of the thesis is reached. Based as it is on a number of new concepts will the proposed method of tailored testing work? The results presented in Chapters 6 and 7, although of interest in giving findings in areas for which no previous research could be found, were principally preparatory for the research reported here.

In attempting to assess how efficient the proposed method is, there are two aspects of its performance that are relevant. One is the number of items needed by a test to reach the termination precision: the other is the ability of the method to converge on to true attainment - and to do so even when misleading initial estimates are provided.

When evaluating the number of items needed the obvious standard of comparison is the number of items in the conventional test being matched in precision. This is the 100-item standard verbal test currently used in recruit allocation. It should be noted, however, that this is not a multiple-choice test. The probability of chance success on the guided-response questions it uses is substantially lower than for the multiple-choice questions of the item library used here. Multiple-choice questions are less efficient gatherers of information because of chance success. And although it is hoped that guessing will be much reduced in live tailored testing it is certainly the case that guessing is a phenomenon present in the response banks used in the simulated testing. Strictly it could be argued that something

like a 125-item 5-option multiple-choice test is the equivalent of the 100-item standard test. This point is mentioned for accuracy of perspective: it is not pressed, as in brief anticipation of the later results it may be mentioned that the 100-item standard is comfortably bettered under all testing conditions, and that under favourable conditions an average of less than forty items proves sufficient.

In checking if the testing procedure converges on true attainment, this has to be defined as the attainment level of the recruits who contributed to any particular response bank. This will be referred to as measured attainment and is an estimate of true attainment subject to a standard error of measurement previously calculated as 1·12 attainment bands (p. 157) - this subject to some possible reduction because of sampling across recruits.

It is a matter of deliberate choice that the proposed procedure presents the outcome of a test as a confidence interval for specified decision risks. In this sense there is a need to establish a convenient comparability between this outcome and the point estimation of the conventional test. This basis is described below. However, in terms of the 90% confidence interval used in reporting the outcome of the simulated tests, barely more than 1% of tests report an interval that does not overlap at least in part with the attainment band (or bands) of the response bank.

The termination precision for the procedure is applied equally for all attainment levels. It is not likely that most conventional tests are equi-precise at all levels, although it is customary for the great majority of test manuals to report only global reliability estimates. In maintaining equi-precision over the attainment range

the procedure may in reality be at some comparative advantage at some attainment levels and at some disadvantage at others. It is because of possible variations in reliability with attainment that decision risks estimated for conventional tests using a standard error of measurement might be expected to be inaccurate.

Finally in this scene-setting Section it is worth emphasising that the simulation being undertaken is based on real-data - answers given by live recruits to specific questions under test conditions. The simulation makes assumptions about the transferability of this response data to another situation, but generally these assumptions are well supported or a worst-case philosophy has applied. The response data contains all the anomalies, errors and guesses that are typical of live testing.

B. Presentation of results

Figure 29 is an illustrative record of the first 28 items of a simulated tailored test. This is the kind of output produced by PROGRAM 5 (Annex XII) which implements the method of Chapter 5.D.

In Figure 29, as the heading details indicate, the response bank for "ABILITY BAND = 10" supplies the recruits' answers for the simulation. This is the wide attainment band 10 response bank introduced earlier. "INITIAL ESTIMATE = 15" shows that on this occasion the program is supplied with misleading information that the (simulated) recruit being tested has a verbal attainment level of about band 15. Consequently the initial upper and lower limits are set as "INITIAL 75% LIMITS 11 to 19", four bands above and below the initial estimate. A series of simulated tests is carried out under this overall condition:

ABILITY BAND= 10 INITIAL ESTIMATE= 15
INITIAL 75% LIMITS 11 TO 19

MAN NO.	ITEM NO.	R-W BALANCE O/ALL	R-W BALANCE LEVEL	MOVE DIRN	TAIL BAND LOCATION SOUGHT R	SOUGHT W	DIFF.	FOUND R	FOUND W	RESP	90 PERCENT LIMITS LOWER	UPPER	DIFF.
1													
	1	-1	-1	0	11		-2	9		0	1.64	14.26	12.62
	2	-2	-1	-1		17	-1		16	0	1.69	14.17	12.48
	3	-3	-1	-1		14	0		14	0	1.56	13.08	11.53
	4	-4	-1	-1		12	0		12	0	1.18	11.34	10.16
	5	-3	1	-1		10	0		10	1	1.84	11.82	9.98
	6	-2	2	0		10	0		10	1	2.91	12.16	9.26
	7	-1	3	0		10	0		10	1	3.80	12.39	8.59
	8	0	4	0		10	0		10	1	4.35	12.60	8.25
	9	1	5	0		10	0		10	1	4.71	12.82	8.11
	10	0	-2	1	11		-2	9		0	4.45	12.12	7.69
	11	1	1	-1		8	0		8	1	4.71	12.30	7.59
	12	0	-3	1	11		-2	9		0	4.50	11.67	7.17
	13	1	2	-1		6	2		8	1	4.71	11.91	7.20
	14	0	-4	1	11		-2	9		0	4.55	11.35	6.82
	15	1	3	-1		6	2		8	1	4.72	11.49	6.77
	16	0	-5	1	11		-2	9		0	4.55	11.14	6.59
	17	1	4	-1		6	2		8	1	4.75	11.29	6.56
	18	0	-6	1	11		-2	9		0	4.57	10.94	6.37
	19	1	5	-1		6	2		8	1	4.74	11.12	6.38
	20	0	-7	1	11		-2	9		0	4.59	10.75	6.16
	21	1	6	-1		6	2		8	1	4.75	10.95	6.20
	22	0	-8	1	11		-2	9		0	4.61	10.55	5.94
	23	1	7	-1		6	2		8	1	4.76	10.78	6.02
	24	0	-9	1	11		-2	9		0	4.63	10.44	5.81
	25	-1	6	-1		6	2		8	0	4.17	9.30	5.13
	26	0	7	0		8	0		8	1	4.37	9.41	5.04
	27	1	8	0		8	0		8	1	4.54	9.53	4.98
	28	2	-8	1	11		-2	9		1	5.08	10.06	4.98
	A	B	C	D	E	F	G	H	I	J	K	L	M

FIGURE 22 An example of a tailored test record.

Figure 29 illustrates part of the first test of this series - testing "MAN NO. 1".

A number of further points will be exemplified by reference to the sequential item record of Figure 29 using the column labels (A) to (M) at the foot. The main feature is that as the item sequence (A) proceeds the 90% confidence limits (K) and (L) converge. (M) gives the width of this confidence interval and testing terminates when this becomes 3·8 or less. (The test in the Figure terminated with item 37 when the interval was from 6·78 to 10·57.)

The record for item 1 shows that a first question was sought (columns E and F) in the R-set of items with a Tail Location in attainment band 11 (E) - the initial lower 75% limit. No such question was available in the response bank and the question found (columns H and I) had a Tail Location in band 9 (H). The question found was the nearest approximation available and is at the harder extremity of the R-set (see Table 8 (p. 188)), but was 2 bands easier (G) than specified. Because of the initial overestimate of the recruit's ability this question is too hard for him and he gets it wrong (shown by a score of 0 in J) - as indeed he does his first four questions while the testing procedure is moving to the appropriate difficulty level. His overall right-wrong balance is now -1 (B), so that for item 2 the procedure specifies a move to an easier question (indicated by -1 in column D of item 2). This is sought in the W-set one step (two attainment bands) below the initial upper limit, that is at band 19-2 = 17 (F).

The procedure then stays within the W-set for a while asking questions at progressively lower bands (I) until right answers are obtained (the first at item 5, column J), and then continues in the

W-set until his overall right-wrong balance is re-established at item 8 (0 in column B). At item 9 a positive balance is established and so for item 10 the procedure switches back to the R-set looking for a harder question (indicated by 1 in D). The procedure then oscillates between the R- and W-sets maintaining a successful overall balance in (B) while the confidence interval for the estimate of the recruit's attainment converges (columns K,L,M).

Annex XV gives, by way of extended illustration, complete tailored test records for 63 (simulated) recruits. These are for three different testing conditions and include an example of one complete test series which continues until the response bank is exhausted. These records are typical of - and constitute a small part of - the simulated tailored tests for which results are reported below. Further details of the annexed records are given at the start of the Annex.

Simulated testing was carried out under seven conditions of response bank and initial estimate. These were given in Table 5 (p. 154) and are repeated in Table 9 for convenience together with the abbreviations to be used in reporting the results.

TABLE 9 The seven attainment conditions used for the simulated testing with their reference abbreviations.

Attainment level by band of the:-		
Response bank	Initial estimate	Condition abbreviation
5	5	5
5	10	5+
10	5	10-
10	10	10
10	15	10+
15	10	15-
15	15	15

Testing under these seven conditions is carried out for both the narrow and wide response banks. The narrow banks are smaller and will not support as lengthy series of simulated tests. Similarly some attainment conditions are more demanding of the response banks than others. (It is the more extreme attainment conditions which are most demanding as these provoke a tail change in the procedure (p. 158) and thereafter draw only upon either the W-set or R-set halves of a bank.)

When a response bank is becoming substantially depleted this is signalled by the testing procedure being no longer able to maintain an overall right/wrong balance with the remaining questions and responses. This triggers a secondary termination criterion built into the testing procedure which can be referred to as an imbalance termination (p. 158). The imbalance termination is entirely a device to signal an objective end to realistic simulations during a test series. Even with the narrow response banks and under the most unfavourable conditions imbalance termination does not occur until well over 600 responses have been withdrawn from a response bank. Imbalance termination is not relevant to live testing and is better regarded as a criterion for simulation termination than for test termination.

When a response bank becomes impoverished and starts to trigger imbalance terminations there is not an immediate dividing line. Instead there is a region in the total test series in which some tests will achieve the required precision before imbalance termination and others will not. Table 10 shows the first occurrence of imbalance termination in the various test series - using the abbreviations of Table 9 to describe the attainment conditions.

TABLE 10 The number of (simulated) recruits tested under various attainment conditions before imbalance termination first occurs.

Attainment Condition	Number of recruits for the given response bank	
	Narrow	Wide
5	22	46
5+	18	46
10-	23	69
10	26	69
10+	-	72
15-	17	50
15	17	53

Based on the values in Table 10 it was determined to analyse 20 recruit test records from each of the narrow bank test simulation series, and to analyse 50 recruit records for each of the wide bank. This gave totals of 140 simulated tests and 350 simulated tests respectively for the narrow and wide response banks. To achieve the even basis of comparison of 20 and 50 simulated tests across the narrow and wide bank conditions meant going beyond the first imbalance termination in some cases. For the narrow banks a total of eight such imbalance terminations were by-passed, and for the wide banks a total of seven. It is not considered that this affects the analysis. The 490 simulated tests analysed below all terminated on reaching the required precision.

To allow the accuracy of convergence of the simulated tests to be

readily reported and compared with the true (measured) attainment of the response banks, the midpoint of the confidence interval at termination will be used. It should be stressed that this is not the intention of the proposed method which attaches considerable importance to reporting upper and lower limits with specified decision risks. The interval midpoint is used only as a convenient basis for comparison; it is an ad hoc compromise of the reporting outcome preferred.

C. Analysis of the simulated tests

1. Test length

Results from the narrow banks and from the wide banks are presented separately in Tables 11 and 12.

The overall pattern in both is the same. Average levels of attainment (band 10) need longer tests to reach the same precision. The higher attainment level (band 15) requires the fewer items of the two non-average levels. In absolute terms the average number of items is well below 100 for all cases, and much less for bands 5 and 15.

The differences between the narrow bank results and those from the wide banks are small: they do not indicate that the somewhat more homogeneous ability in the narrow banks permits shorter tests. Consequently attention will be directed to the larger samples of the wide banks.

Within any one attainment level - 5, 10, or 15 - the differences in Table 12 between the mean numbers of items are small. Now, all the tests within an attainment level are from the same response bank and so the test samples for the two or three conditions within a level

TABLE 11 The test lengths needed to reach the required precision for twenty tests under various attainment conditions using narrow response banks.

	Attainment Conditions						
	5	5+	10-	10	10+	15-	15
No. of items[a]							
5							
10							1
15							0
20						2	1
25						2	4
30		2				2	1
35	3	3	1	1	1	3	3
40	3	2	2	1	1	4	3
45	5	2	0	1	0	1	1
50	1	5	1	1	3	4	4
55	3	1	3	3	3	2	1
60	1	2	1	1	1		1
65	1	0	2	2	0		
70	2	1	1	3	2		
75	1	0	1	2	2		
80		2	1	1	2		
85			2	2	2		
90			1	0	0		
95			2	0	0		
100			0	0	1		
103+[b]	—	—	2	2	2	—	—
Total:	20	20	20	20	20	20	20
Mean Length[c]:	50.1	49·1	70·9	67·4	70·4	38·4	38·0
Standard Deviation[c]:	12·5	14·6	22·4	19·8	23·7	10·8	12·8

Notes:

a The nominal values include lengths down to two less and up to two more than the given length.

b The lengths of the six tests in this interval were 103, 103, 106, 114, 115, and 131 items.

c Calculated from the exact values before grouping.

TABLE 12 The test lengths needed to reach the required precision for fifty tests under various attainment conditions using wide response banks.

	Attainment Conditions						
	5	5+	10-	10	10+	15-	15
No. of Items[a]							
5							
10							1
15						1	2
20						3	8
25	2	3				8	5
30	3	5		1	3	7	4
35	5	4	1	2	1	5	7
40	8	3	2	1	3	5	4
45	6	9	3	8	6	8	5
50	1	8	7	3	1	5	5
55	9	3	7	4	6	3	4
60	4	4	11	6	5	3	2
65	2	3	7	8	6	1	1
70	3	4	2	4	7	0	1
75	2	0	3	5	0	1	1
80	1	2	2	0	3		
85	1	0	1	3	3		
90	2	2	0	1	1		
95	0		1	2	1		
100	0		0	0	1		
103+[b]	1		3	2	3		
Total	50	50	50	50	50	50	50
Mean[c]	53·0	50·3	62·6	64·4	63·6	38·9	37·9
S.D.[c]	19·0	15·9	16·7	21·0	20·5	13·5	15·2

Notes:

a As Table 11.

b The lengths of the nine tests in this interval were 120, 104, 104, 108, 109, 117, 118, 124, and 143.

c Calculated from the exact values before grouping.

cannot be held to be independent. Hence the sampling error of the difference between condition means within attainment level is probably less than the assumption of independence would indicate. Also all the distributions are skewed towards longer tests: on this count the standard deviations are raised considerably by the few long tests. Putting these two situations together it can be held that using the lowest variance estimate within an attainment condition to estimate the sampling error of the differences in means will still constitute a conservative procedure.

None of the within attainment mean differences give rise to student's t values approaching significance.

The comparisons across the attainment levels use different response banks and the samples are independent. Taking the smallest variance estimate within each attainment level gives $13{\cdot}5^2$ as the smallest overall variance and $16{\cdot}7^2$ as the largest. The variance ratio of 1·53 is not significant at the 5% level. A pooled estimate of the variance is $15{\cdot}58^2$, giving a sampling error of 3·12 ($= SD\sqrt{1/n_1 + 1/n_2}$). The difference between the closest means from the 5 and 10 attainment levels is 17·3 (= 67·4 - 50·1), and between the 5 and 15 bands is 10·7 (= 49·1 - 38·4). Both differences give t values in excess of three and significant beyond the 1% level.

The above tests are all conservative. The differences in test lengths between attainment levels are well established. Differences caused by misleading initial estimates within attainment levels are not established: additionally, in comparing the narrow and wide banks for attainment level 10, the differences in means are not consistent. It is unlikely that there are differences of any

practical significance caused by the initial misestimates.

2. Accuracy of convergence

Tables 13, 14 and 15 present the distributions of the termination interval midpoints for the three attainment levels (5, 10 and 15) respectively for the various attainment conditions and for the narrow and wide response banks. This midpoint is the ad hoc compromise described in Section A of this Chapter to be used as a convenient basis for comparison.

In general terms the midpoints are all close to the measured attainments of their response banks. The largest deviations occur for attainment level 5 where the largest difference is 0·76. At attainment level 10 the largest difference is -0·28, and at level 15 it is -0·14. Most of the remaining differences are in the same directions but much smaller. One factor at work here may be a regression to the mean consequent upon the original selection of recruits for particular attainment levels on a particular test administration. This would fit with the direction of the difference for the 5 and 15 attainment levels. It would not explain the progressive decrease in the difference with increasing attainment unless increasing reliability with attainment was assumed also. Guessing will systematically decline with increasing attainment, but it is hard to see how guessing can have any residual effect not already taken into account by the inclusion of guessed right answers in the item conditional probabilities.

In any case the larger deviation from measured attainment at attainment level 5 is accompanied by a smaller standard deviation. There is apparently a slip in location accompanied by greater stability. This could be suggestive of an inability of the item library to locate

TABLE 13 The distribution of the termination interval midpoints for the narrow and wide response banks at attainment level 5.

Attainment band of midpoint[a]	Narrow response bank Attainment condition: 5	Narrow response bank Attainment condition: 5+	Wide response bank Attainment condition: 5	Wide response bank Attainment condition: 5+
7·5 - 8	1	1		
7 - 7·5	1	1		1
6·5 - 7	0	2	1	0
6 - 6·5	4	2	9	4
5·5 - 6	3	4	8	14
5 - 5·5	7	7	13	13
4·5 - 5	4	2	9	8
4 - 4·5		1	6	7
3·5 - 4			4	3
3 - 3·5	—	—	—	—
Total	20	20	50	50
Mean[b]	5·68	5·76	5·23	5·19
S.D.[b]	0·846	0·877	0·813	0·669

Notes:

a The intervals are exclusive of the upper limit.

b Calculated on the exact values before grouping.

TABLE 14 The distribution of the termination interval midpoints for the narrow and wide response banks at attainment level 10.

Attainment band of midpoint[a]	Narrow response bank Attainment condition: 10-	10	10+	Wide response bank Attainment condition: 10-	10	10+
13 - 13·5					1	
12·5 - 13				1	2	
12 - 12·5			1	2	1	3
11·5 - 12	2	1	1	0	4	4
11 - 11·5	0	2	1	3	2	3
10·5 - 11	3	2	3	5	6	2
10 - 10·5	2	1	2	8	4	5
9·5 - 10	2	7	5	11	8	13
9 - 9·5	7	3	1	9	9	10
8·5 - 9	3	2	3	5	7	4
8 - 8·5	1	1	3	4	6	4
7·5 - 8		1		2		0
7 - 7·5						2
Total	20	20	20	50	50	50
Mean[b]	9·75	9·72	9·81	9·79	9·95	9·84
S.D.[b]	0·942	0·979	1·156	1·120	1·282	1·196

Notes:

a The intervals are exclusive of the upper limit.

b Calculated on the exact values before grouping.

TABLE 15 The distribution of the termination interval midpoints for the narrow and wide response banks at attainment level 15.

Attainment band of midpoint[a]	Narrow response bank Attainment condition: 15-	Narrow response bank Attainment condition: 15	Wide response bank Attainment condition: 15-	Wide response bank Attainment condition: 15
17 - 17·5				2
16·5 - 17	1	1	2	0
16 - 16·5	2	0	3	8
15·5 - 16	3	3	9	6
15 - 15·5	3	5	13	12
14·5 - 15	3	3	6	5
14 - 14·5	4	6	9	7
13·5 - 14	2	1	2	4
13 - 13·5	1	1	3	1
12·5 - 13	1		3	5
12 - 12·5	—	—	—	—
Total	20	20	50	50
Mean[b]	14·93	14·86	14·92	14·97
S.D.[b]	1·048	0·772	0·978	1·175

Notes:

a The intervals are exclusive of the upper limit.

b Calculated on the exact values before grouping.

testees at the lowest attainment levels - possibly because of some deficiency in the upper tails of the question curves at this level.

The differences from measured attainment are sufficiently small to give confidence in the ability of the procedure to converge with considerable accuracy. The midpoint estimator is an ad hoc device of convenience never intended to do more than provide the broad basis for the comparison being reported.

In conjunction with the accuracy of convergence the small standard deviations are also impressive of the adequacy of the procedure. It will be recalled that the standard error of measurement of the standard verbal test used for allocating recruits to response banks is 1·12 attainment bands. This, or less, is generally the size of the standard deviation found. It is possible for the standard deviation to be a little less because of the refinement of sampling across recruits. Given that the parallel forms reliability of the standard verbal test is 0·94, its correlation with true attainment may be estimated as $\sqrt{0{\cdot}94}$, that is 0·969. This level of reliability would give a standard error of measurement of 0·8. The observed values can be readily construed as representative of population values lying in the interval 0·8 to 1·12.

The differences in means within attainment conditions do not consistently favour the accurate initial estimate, although there is some tendency for this to be so. The differences are all extremely small - 0·16 is the largest - so that the procedure does seem practically resistant to being misled by poor estimates.

The differences between narrow and wide banks are small - usually

favouring the wide bank rather than the narrow in terms of proximity to measured attainment. Regression to the mean would have most effect on a more narrowly selected group.

D. Summary and additional comments

In round terms the procedure has been able to achieve the required precision of measurement in an average of 40, 50, or 70 items depending on the attainment level. This precision has on the average been within half a band or less of the measured attainment as judged by the mid-point of the termination interval. Judged more in the spirit of the procedure's focus on decision risks, 485 of the 490 simulated recruits completed their tests with a termination interval that overlapped with the attainment level of the response bank. The number falls only to 469 (96%) even if the attainment level of a wide response bank is defined rigorously as its central band. The stability, too, of the test estimates is fully comparable with the standard verbal test on which the response banks are based.

A step size of two attainment bands was used here when the procedure made changes in the Tail Location of its next-question specification. In a small number of runs a step size of two bands rather than one band had shown a little advantage in dealing with initial misestimates.

For individual (simulated) recruits 15 out of the 490 took tests longer than 102 items. Inspection of this 3% of records showed that convergence had been extremely slow for the latter part of these tests. Little would have been lost by a substantially earlier termination. This suggests that progress to termination might be a useful concept in picking out such recruits rather than simply imposing a maximum test

length. These cases could then be interrupted and difficulty changes introduced, or testing restarted to break out of the groove. Such cases might be largely simulation phenomena.

9. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH.

The results obtained by the method of tailored testing developed here encourage acceptance of the utility of some at least of the concepts on which it is based. However, while the procedure has achieved accuracy and precision with shorter tests, what it is not possible to investigate with real-data simulation is the time element. Will testees take as long to answer the smaller number of questions as is allowed for the administration of a conventional test? If a degree of speeding is introduced into tailored testing in an effort to translate the savings in test length into time savings, what effect will this have on the accuracy of convergence?

These questions open up large areas of research which can only be investigated by empirical studies which seat recruits at online terminals. In conventional testing it is difficult to obtain information about the time taken by testees on individual items. This is an area on which it might be said there is a considerable body of ignorance. This area is a part of the larger one referred to and concerned with the interaction of attainment (and possibly other characteristics) and item timings. Item timings could well be considered as additional criteria in item analysis.

More specifically on the testing procedure as it stands, there are many details in which the present procedure no doubt fails to implement its concepts optimally. This is inevitably so for a new proposal. Study of the aspects of item performance that contribute to convergence will probably quickly replace the simple ideas of Tail Location and Tail Discrimination with more effective indices. As an immediate step, test length, from a simulation repeatedly using a single item across

testees for different attainment levels, would perhaps be a useful criterion for identifying effective items.

The advantage that the simulation has over the live testing immediately possible is that it is able to pretend that one item is several identical items. Although the average test lengths indicate that an item library of modest size will probably be sufficient, further items of the kind at present available may be helpful or even necessary in achieving live testing results comparable with those of the simulation. This is an area which could be profitably researched by further real-data simulation studies. If access to the response bank were to be restricted in selected ways the effect of item library deficiences on test efficiency could be determined.

In more applied terms, considering the implementation of an operational system, the ongoing calibration of the item library must be provided for. The items in the simulation item library are calibrated against a conventional test. Will group testing for calibration need to be retained as an occasional exercise? Or can a tailored testing system maintain its own internal consistency? Systematic methods are needed to do this.

Two other background system requirements of importance are the capacity to introduce new questions, and monitoring of item performance. New questions can perhaps be introduced automatically as one or two unscored questions within a tailored test. In this way information would be obtained about their performance until either the questions would be added to the library or discarded. Old questions that monitoring detects as having deteriorated in performance would also be

automatically deleted from the library.

At a selection centre tailored testing should eventually extend to the range of psychological characteristics measured rather than simply the choice of questions for a given characteristic. The quintessential quality of tailored testing is its individual flexibility. If a selector would be helped by information from an area susceptible to testing this ideally would be provided by tailored testing sessions integrated with ongoing guidance and briefing activities. Either a retest on an attainment which becomes more critical during later interviewing, or a test of some specialised ability could be individually available without administrative formality.

This kind of idealised selection centre is a long way from the small step taken by the present research. Tailored testing will have to offer initial technical advantages even if eventually its major benefit is a humanitarian respect for individuality. The proposed testing procedure apparently offers some technical gain - and the further research areas sketched out may be able to consolidate this and advance beyond.

10. REFERENCES

Anastasi, A. An empirical study of the applicability of sequential analysis to item selection. Educational and Psychological Measurement, 1953, 13, 3-13.

Anastasi, A. Psychological Testing. New York: Macmillan, 1954.

Anderson, T.W., McCarthy, P.J., & Tukey, J.W., "Staircase" methods of sensitivity testing. Navord Report 65-46. Statistical Research Group, Princeton University, 1946, 1-134.

Angoff, W.H., & Huddleston, E.M., The multi-level experiment. A study of a two-level test system for the College Board Scholastic Aptitude Test. Statistical Report 58-21, Princeton, New Jersey: Educational Testing Service, 1958.

Armitage, P. Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. Journal of the Royal Statistical Society, 1950, 12, 137-144.

Atkinson, R.C., & Wilson H.A. (Eds.) Computer Assisted Instruction. New York: Academic Press, 1969.

Bayroff, A.G., Thomas, J.J., & Anderson, A.A. Construction of an experimental sequential item test. Research Memorandum 60-1 Personnel Research Branch, US Dept. of Army, 1960.

Bayroff, A.G. Feasibility of a programed testing machine. Report 64-3, US Army Personnel Research Office, 1964.

Bayroff, A.G., & Seeley, L.C. An exploratory study of branching tests. Technical Research Note 188, Washington D.C. US Army BESRL, 1967.

Bayroff, A.G., Ross, R.M., & Fischl, M.A. Development of a programed testing system. Technical Paper 259, US Army, RIBSS, 1974.

Betz, N.E., & Weiss, D.J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, University of Minnesota, 1973.

Betz, N.E. & Weiss, D.J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, University of Minnesota, Minneapolis, 1974.

Betz, N.E. & Weiss, D.J. Empirical and simulation studies of flexilevel ability testing. Research Report 75-3, Psychometric Methods Program, University of Minnesota, 1975.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord & Novick (Eds.), 1968, Chapters 17-20.

Brownlee, K.A., Hodges, J.L., & Rosenblatt, M. The up-and-down method with small samples. Journal of the American Statistical Association, 1953, 48, 262-277.

Bryson, Rebecca. A comparison of four methods of selecting items for computer assisted testing. Technical Bulletin STB 72-8, San Diego, California, Psychological Sciences Division, ONR, 1971.

Bryson, Rebecca. Shortening Tests: effects of method used, length, and interval consistency on correlation with total score. Proceedings, 80th Annual Convention of the American Psychological Association, Honolulu, 1972, 7, 7-8.

Burgess, G.C. Use of sequential analysis for determining test item difficulty level. Educational and Psychological Measurement, 1955, 15, 80-86.

Cleary, T.A. Linn, R.L., & Rock, D.A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-60, a.

Cleary, T.A., Linn, R.L., & Rock, D.A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187, b.

Cochran, W.G., & Davis, M. Stochastic approximation to the median effective dose in bioassay. In J. Gurland (Ed.), Stochastic models in Medicine and biology, Madison: University of Wisconsin Press, 1964, pp. 281-300.

Cochran, W.G., & Davis, M. The Robbins-Monro method for estimating the median lethal dose. Journal of the Royal Statistical Society, Series B, 1965, 27, 28-44.

Cornsweet, T.N. The staircase method in psychophysics. American Journal of Psychology, 1962, 75, 485-491.

Cowden, D.J. An application of sequential sampling to testing students. Journal of American Statistical Association, 1946, 41, 547-556.

Cronbach, L.J., & Gleser, G.C. Psychological tests and Personnel decisions. Urbana, University of Illinois Press, 1957 (1st. ed.) 1965 (2nd. ed.).

Davis, M. Comparison of sequential bioassays in small samples. Journal of the Royal Statistical Society, Series B, 1971, 33, 78-87.

Dixon, W.J., & Mood, A.M. A method for obtaining and analyzing sensitivity data. Journal of American Statistical Association, 1948, 43, 109-126.

DuBois, P.H. Varieties of psychological test homogeneity. American Psychologist, 1970, 25, 532-536.

Duncan, K.D. Experiments with an inexpensive device for programmed instruction in the multiple choice branching style. Programmed Learning, 1, 145-154, 1964.

Feldt, L.S., & Forsyth, R.A. An examination of the context effect in item sampling. Journal of Educational Measurement, 1974, 11, 73-82.

Ferguson, R.L. Computer-assisted criterion-referenced testing. Working paper No.49, University of Pittsburgh, Learning Research & Development Center, 1969.

Ferguson, R.L. Computer assistance for individualising measurement. Report 1971/8, Pittsburgh, Pa., University of Pittsburgh, Learning Research and Development Center, 1971, a.

Ferguson, R.L., & Hsu, T. The application of item generators for individualising measurement. Report 1971/14, University of Pittsburgh, Learning Research and Development Center, 1971, b.

Flaugher, R.L., Melton, R.S., & Myers, C.T. Item re-arrangement under typical test conditions. Educational and Psychological Measurement, 1968, 28, 813-824.

Freeman, P.R. Optimal Bayesian sequential estimation of the median effective dose. Biometrika, 1970, 57, 79-89.

Green, B.F. Comments on tailored testing. In W.J. Holtzman (Ed.), 1970.

Greenwood, D.I., & Taylor, C. Adaptive testing in an older population. Journal of Psychology, 1965, 60, 193-198.

Hansen, D.N. An investigation of computer-based science testing. In Atkinson, R.C., & Wilson, H.A. (Eds.), 1969.

Hansen, D.N., Johnson, B.F., Fagan, R.L., Tam, P., & Dick, W. Computer-based adaptive testing models for the air force technical training environment Phase 1: Development of a computerized measurement system for air force technical training. Report AFHRL-TR-74-48, Air Force Human Resources Laboratory, Technical Training Division, Lowry Air Force Base, Colorado, 1974.

Hick, W.E. Information theory and intelligence tests. British Journal of Psychology, Statistics Section, 1951, 4, 157-164.

Holtzman, W.H. (Ed.) Computer-assisted instruction, testing and guidance, New York: Harper and Row, 1970.

Huck, S.W., & Bowers, N.D. Item difficulty level and sequence effects in multiple-choice achievement tests. Journal of Educational Measurement, 1972, 9, 105-111.

Hutt, M.L. A clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet. Journal of Consulting Psychology, 1947, 11, 93-103.

Jensema, C.J. An application of latent trait mental test theory. British Journal of Mathematical and Statistical Psychology, 1974, 27, 29-48.

Kent, G.H. Suggestions for the next revision of the Binet-Simon scale. Psychological Record, 1937, 409-432.

Killcross, M.C., & Cassie, A. The potential use of tailored testing for allocation to Army employments. Occasional Note APRE 41/73, Farnborough, Hants: Army Personnel Research Establishment, 1973. also in Singleton, W.T., & Spurgeon, P. (Eds.) Measurement of human resources, London: Taylor & Francis, 1975, pp. 117-122.

Killcross, M.C. A tailored testing system for selection and allocation in the British Army. Montreal: paper presented at the 18th International Congress of Applied Psychology, 1974.

Killcross, M.C., Hammond, D.R.F., & Preston, L.R. Revision of the selection centre test battery: I. A multiple-choice verbal test. APRE Report 41/75, Farnborough, Hants: Army Personnel Research Establishment, (in press).

Krathwohl, D.R., & Huyser, R.J. The sequential item test. American Psychologist, 1956, 11, 419 (Abstract).

Keston, H. Accelerated stochastic approximation. Annals of Mathematic Statistics, 1958, 29, 41-59.

Larkin, K.C., & Weiss, D.J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3, Psychometric Methods Program, University of Minnesota, 1974

Larkin, K.C., & Weiss, D.J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 75-1 Psychometric Methods Program, University of Minnesota, 1975.

Lazarsfield, P.F. Latent structure analysis. In S. Koch (Ed.) Psychology: a study of a science. Vol. 3, New York: McGraw-Hill, 1959, 476-542.

Linn, R.L., Rock, D.A., & Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Linn, R.L., Rock, D.A., & Cleary, T.A. Sequential Testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F.M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F.M. Item sampling in test theory and in research design. Research Bulletin 65-22, Princeton, New Jersey: Educational Testing Service, 1965.

Lord, F.M. Item characteristic curves estimated without knowledge of their mathematical form - a confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50, a.

Lord, F.M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), 1970, b.

Lord, F.M. Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31, a.

Lord, F.M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151, b.

Lord, F.M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-242, c.

Lord, F.M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813, d.

Lord, F. M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711. e.

Lord, F.M. Individualised testing and item characteristic curve theory. also in Atkinson R.C. et al (Eds.), Contemporary developments in mathematical psychology, Vol. 2, San Francisco: W.H. Freeman & Co, 1974, a.

Lord, F.M. Practical methods for redesigning a homogeneous test, also for designing a multilevel test. Research Bulletin 74-30, Educational Testing Service, Princeton, New Jersey, 1974, b.

Lord, F.M. The 'ability' scale in item characteristic curve theory. Psychometrika, 40, 1975, 205-217, a.

Lord, F.M. A broad-range tailored test of verbal ability. Research Bulletin 75-5, Princeton, New Jersey, Educational Testing Service, 1975, b.

Lord, F.M., & Novick, M.R. (Eds.) Statistical Theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Marso, R.N. Test item arrangement, testing time, and performance. Journal of Educational Measurement, 1970, 7, 113-118.

McBride, J.R., & Weiss, D.J. Recent and projected developments in ability testing by computer. Paper presented at "Occupational Research and the Navy: Prospectus 1980." a symposium sponsored by the Navy Personnel Research and Development Center, San Diego, California, 1973.

McBride, J.R., & Weiss, D.J. A word knowledge item pool for adaptive ability measurement. Research Report 74-2, Psychometric Methods Program, University of Minnesota, 1974.

McCarthy, P.J. A class of methods for estimating reaction to stimuli of varying severity. Journal of Educational Psychology, 1949, 40, 143-156.

McGill, P.A. The concept of a programmed, branching, or sequential item test. Paper presented at the Defence Psychologists Symposium, Shrivenham, 1968. Army Personnel Research Establishment, 1968.

McNemar, Q. The revision of the Stanford-Binet Scale. Boston: Houghton Mifflin Co., 1942.

Mollenkopf, W.G. An experimental study of the effects on item-analysis data of changing item placement and test time limit. Psychometrika, 1950, 15, 297-315.

Moonan, W.J. Some empirical aspects of the sequential analysis technique as applied to an achievement examination. Journal of Experimental Education, 1950, 18, 195-207.

Mussio, J.J. A modification to Lord's model for tailored tests. Unpublished Doctoral dissertation, University of Toronto, 1972.

Owen, R.J. A Bayesian approach to tailored testing. Research Bulletin 69-92, Princeton, New Jersey, Educational Testing Service, 1969.

Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Owens, T.R., & Stufflebeam, D.L. An experimental comparison of item sampling and examinee sampling for estimating test norms. Journal of Educational Measurement, 1969, 6, 75-83.

Patterson J.J. An evaluation of the sequential method of psychological testing. Doctoral dissertation, Michigan State University, 1962. also Ann Arbor, Michigan: University Microfilms, 1962, No. 63-1748.

Reckase, M.D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behaviour Research Methods & Instrumentation, 1974, 6, 208-212, a.

Reckase, M.D. An application of the Rasch simple logistic model to tailored testing. Paper presented at the Annual Meeting of the American Educational Research Association, 1974, b.

Robbins, H., & Monro, S. A stochastic approximation method. The Annals of Mathematical Statistics, 1951, 22, 400-407.

Rose, R.M., Teller, D.Y., & Rendleman, P. Statistical properties of staircase estimates. Perception and Psychophysics, 1970, 8, 199-204.

Sax, G., & Cromack, T.R. The effects of various forms of item arrangement on test performance. Journal of Educational Measurement, 1966, 3, 309-311.

Seeley, L.C., Morton, M.A., & Anderson, A.A. Exploratory study of a sequential item test. Technical Research Note 129, Washington D.C. US Army Personnel Research Office, 1962.

Sirotnik, K. An investigation of the context effect in matrix sampling. Journal of Educational Measurement, 1970, 7, 199-207.

Statistical Research Group, Columbia University. Sequential analysis of statistical data: applications. New York: Columbia University Press, 1945.

Stocking, M. Short tailored tests. Research Bulletin 69-73, Princeton, New Jersey, Educational Testing Service, 1969.

Taylor, M.M., & Creelman, C.B. PEST: efficiency estimates on probability functions. Journal of the Acoustical Society of America, 1967, 41, 782-787.

Terman, L.M. in McNemar, Q., 1942, Chapter 1.

Tsutakawa, R.K. Asymptotic properties of the block up-and-down method in bioassay. The Annals of Mathematical Statistics, 1967, 38, 1822-1828.

Urry, V.W. A Monte Carlo investigation of logistic mental test models. Unpublished Doctoral dissertation. Purdue University, 1970.

Urry, V.W. Individualised testing by Bayesian estimation. Bureau of Testing, University of Washington, April 1971, a.

Urry, V.W. Approximation methods for the item parameters of mental test models. Bureau of Testing, University of Washington, December 1971, b.

Wald, A. Sequential analysis. New York: Wiley, 1947.

Wald, A. Statistical decision functions. New York: Wiley, 1950.

Waters, C.J. Preliminary evaluation of simulated branching tests. Technical Research Note 140, Washington, D.C. US Army Personnel Research Office, 1964.

Waters, C.W., & Bayroff, A.G. A comparison of computer-simulated conventional and branching tests. Educational and Psychological Measurement, 1971, 31, 125-136.

Weiss, D.J., & Betz, N.E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, University of Minnesota, 1973, a.

Weiss, D.J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, University of Minnesota, 1973, b.

Weiss, D.J. Strategies of adaptive ability measurement. Research Report 74-5, Minnesota, MN. Psychometric Methods Program, University of Minnesota, 1974.

Wetherill, G.B. Sequential estimation of quantal response curves. Journal of the Royal Statistical Society, Series B, 1963, 25, 1-38.

Williams, E.J. Experimental designs balanced for the estimation of residual effects of treatments. Australian Journal of Scientific Research, 1949, 3A, 351-363.

Wood, R. The efficacy of tailored testing. Educational Research, 1969, 11, 219-222.

Wood, R, Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.